E BELL JOURNAL OF ECONOMICS
D MANAGEMENT SCIENCE

l. 3 . 1972

2

Editor:
Paul W. MacAvoy, *Massachusetts Institute of Technology*

Associate Editors:
Dale W. Jorgenson, *Harvard University*
Harvey J McMains, *American Telephone and Telegraph Company*
William F. Sharpe, *Stanford University*

Business Manager:
Phillip E. Davis, *American Telephone and Telegraph Company*

CUL -H05426-2- P5025

RING '72
0 1

Table of contents

ARTICLES

3 **Profit limitation: regulated industries and the defense-space industries**
ARTHUR E. BURNS

26 **Cost allocation in railroad regulation**
ZVI GRILICHES

42 **Antitrust laws and public utility regulation**
PHILLIP E. AREEDA

58 **The application of finance theory to public utility rate cases**
STEWART C. MYERS

98 **The appropriate scope of regulation in the cable television industry**
RICHARD A. POSNER

130 **Prospects for cable in the 100 largest television markets**
ROLLA EDWARD PARK

151 **Corporate investment under uncertainty and Pareto optimality in the capital markets**
MICHAEL C. JENSEN AND JOHN B. LONG, JR.

175 **Telephone communications in Canada: demand, production, and investment decisions**
A. RODNEY DOBELL, LESTER D. TAYLOR, LEONARD WAVERMAN, TSUANG-HUA LIU, AND MICHAEL D. G. COPELAND

220 **Investment behavior and the production function**
DALE W. JORGENSON

252 **Pricing pollution and other negative externalities**
RICHARD A. TYBOUT

267 **Models for determining least-cost investments in electricity supply**
DENNIS ANDERSON

CALCUTTA UNIVERSITY CENTRAL LIBRARY

P-5025

Table of contents (cont'd)

COMMENTS AND REVIEWS

302 **Operations research and systems analysis in the public sector**
PAUL R. KLEINDORFER

307 **The theory of peak-load pricing: a final note**
GEORGE W. WILSON

311 **Note on "gold plating" or "rate base padding"**
EDWARD E. ZAJAC

316 **Cash-flow patterns and the choice of a depreciation method**
ALAN KRAUS AND RONALD J. HUEFNER

335 **Metering costs and marginal cost pricing in public utilities**
G. FRANKLIN MATHEWSON AND G. DAVID QUIRIN

340 **A state preference approach to public utility pricing and investment under risk**
STEPHEN C. LITTLECHILD

346 **User charges as a means for pollution control: the case of sewer surcharges**
DON ETHRIDGE

354 **Errata**

Profit limitation: regulated industries and the defense-space industries

Arthur E. Burns
Professor of Economics
George Washington University

An effective profit limitation policy must employ control techniques that yield a profit outcome, or ceiling, consistent with policy norms. In general, utility profit regulation achieves this by allowing a profit sufficient to attract capital and provide a "fair" return including a risk premium. Whatever else may be said, such regulation is articulated.

In the defense-space industries fee determination is generally cost-based and a minor element in procurement price policy. Profit outcomes are indeterminant, varying with capital turnover rates. Subsequent renegotiation provides for the recapture of "excessive" profits. But the Renegotiation Act establishes no norms; the Act is devoid of analytical guidelines. There is no articulated profit limitation policy in the defense-space sector.

Studies reveal "inequities" in cost-based fee arrangements and progress payment policy. Prime contractors fare better than subcontractors, as do large compared with small contractors. Capital intensity and cost reduction are discouraged. To lessen these deficiencies a 1972 landmark policy shift gives capital equal weight with cost in a partial acceptance of regulated profit policy.

This paper suggests a functional profit analysis as a renegotiation policy guide, based on court decisions and Department of Defense policy change.

1. Introduction

■ The policy of governmental profit limitation prevails generally in the regulated industry sector and in the wide range of businesses in the defense-space procurement area. In both sectors the avowed purpose is to achieve "equity," in some sense. Price (rate) setting and, to a lesser extent, profit or revenue refunds are the means used by the profit control bodies to attain this end. The similarities end at this point. The regulated industries are subject to a controlled profit outcome, the rate of which varies within a fairly narrow range. Such control is absent in the defense-space sector of the economy; there is no defined profit outcome sought by the control agencies. The object is to secure a "fair" price for the government and a "fair"

Arthur E. Burns obtained the A.B. and M.A. degrees from the University of California, Berkeley, and the Ph.D. from George Washington University, all in economics. He has been a consultant to the International Cooperation Administration, the White House, and the U. S.-Puerto Rico Status Commission. In 1962 he served as an expert witness in the North American Aviation renegotiation case before the Tax Court and in subsequent years was consultant to the Renegotiation Board. He is Dean of the Graduate School of Arts and Sciences at George Washington University.

The author is grateful to Professor Ralph C. Nash, George Washington University Law School, for his helpful comments on this paper.

profit for the contractors, and to recapture "excessive" profits when such are determined to exist. The key element in this triad is the meaning and measurement of profit. It is here that the two sectors under discussion show wide divergences.

The scope of this paper is limited to a few central questions basic to the policy aspects of profit limitation in these two sectors:

(1) What is the meaning of the profits allowed business by the control bodies? What services or functions are compensated by the allowable profits? Implicitly or explicitly, this question requires an answer to guide reasonably consistent policy.
(2) What is the base against which the dollars of profit are measured? Without some such measure—a profit ratio—profit control policy lacks moorings. The answer to this question rests upon the meaning of profitability.
(3) What are the criteria, or norms, that guide the determination of the allowable profit? This entails comparisons of the controlled firms and sectors with other sectors, all of which draw on the pool of economic resources having alternative opportunities.
(4) What are the techniques used to restrict profit and how are these techniques related to the desired profit outcome?

These questions bring out points of similarity and contrast; the contrasts dominate the picture. The defense-space sector is not a regulated industry in the accepted meaning of this term. Indeed, its vast diversity in financial structure, output, size, and production complexity precludes the degree of profit control found in the regulated industries.

Most of this paper dwells on profit limitation in the defense-space sector, with appropriate references to the regulated industries. There are two phases of profit limitation in the defense-space sector, each wholly independent.

(1) The first phase is the negotiation of individual contracts by the procurement agencies and the contractors. At this point the fee (profit) on each contract is established *ex ante*, usually as a component of the total estimated cost (price) of the contract.[1] In some instances terms are set forth for the subsequent revision of the contract price and therefore the amount of the fee. Disputes that arise at contract settlement over allowable costs and fees are usually settled by negotiation. In some cases this fails and the parties go to the Armed Forces Board of Contract Appeal, and in rare cases to the United States Court of Claims.
(2) The second phase is profit renegotiation, an *ex post* review of a contractor's total profit (received or accrued) from all contracts during his fiscal year. This is the responsibility of the Renegotiation Board, which is independent of the procurement agencies. When the Board finds "excessive" profits in this review, it re-

---

[1] The terminology may need clarification. Contract negotiators and contract lawyers use the term "fee" only for cost reimbursable fixed-fee contracts—about 30 percent of the negotiated procurement dollars. Fees negotiated on other contracts are called "profits." In this paper I call all negotiated contractor remuneration a "fee" and reserve the term "profit" to the profit outcome or residual. On the average, the latter is about one-half or less of the initially negotiated remuneration.

captures them, either by agreement or by order. Contractors may petition the United States Court of Claims for a redetermination of the issue if they feel grossly aggrieved by a Board determination.[2]

Before reviewing these profit control arrangements a few positions need stating. A profit limitation policy must satisfy the requirement that the control techniques result in a profit outcome for the firms in the controlled group, consistent with policy objective norms. The norm or norms must be quantitative and measured against some base. Within limits, policy sets forth what profit is allowable and what is not. Allowability implies some conception of what it is that profit compensates; otherwise the distinction is capricious.

In general, the regulated industries satisfy these requirements: The approved rate or rates, including the margin over costs, are intended to yield a profit outcome measured against the capital base which approximates the regulating authority norm—a profit rate sufficient to attract needed capital and provide a "fair return" to investors, i.e., essentially a capital yield plus a risk element. Stating it this way implies no justification for such profit control policy; in some areas the control may be needless, or too costly and restrictive, or a sheltering device. The critical literature is replete with such views.[3]

Be that as it may, it is an articulated policy. This cannot be said of profit limitation efforts in the defense-space industries. At the procurement phase, fee policy is, for the most part, a minor element in procurement price policy, with an indeterminate profit outcome and no clearly defined profit norms.[4] The Renegotiation Act provides for the recapture of "excessive" profits but establishes no norms for its determination beyond a list of elements the Renegotiation Board must "take into account."[5] The Act is simply devoid of analytical guidelines. This is not to suggest a regulated industry approach for the government procurement market, but as will be shown later, some aspects of it may be appropriately incorporated into a policy of broader scope.

2. Defense contract profit limitations

■ The cost basis for negotiating defense contract fees came into prominence during World War I, partly replacing the traditional and statutory competitive bid procedure.[6] Some contracts were cost-plus-a-percentage-of-cost (CPPC) and others were cost-plus-fixed-fee (CPFF), some of which had an incentive feature providing as much as 50–50 sharing by contractors of actual cost underruns below the initial estimated cost. Charges of "profiteering" soon arose and grew in volume, leading to government efforts to recapture part of the profits by withholding contract settlement payments and court action. Later, in 1934, Congress passed the Vinson-Trammell Act,[7] placing a profit ceiling of 10 percent of the cost of naval vessel construction and 12 percent of aircraft costs, with any excess refundable to the Treasury. With the approach of World War II, other, and more stringent, profit limitation bills were proposed in the Congress, again with profit ceilings related to cost.

---

[2] Prior to the mid-1971 extension of the Renegotiation Act of 1951 [13], the Tax Court of the United States had jurisdiction over these cases. Later in this article some Tax Court decisions will be discussed briefly.

[3] For example, see Stigler and Friedland [9], Posner [8], and MacAvoy [5,6]. The list is long.

[4] As will be shown later, recent Department of Defence (DOD) policy changes are moving in the direction of an articulated profit policy.

[5] See [13].

[6] Procurement by formal advertising for competitive bids was required by statute as early as 1809. Because of massive procurements, technological complexities, and the need for speedy action, exceptions were permitted on a substantial scale in World War I for the first time. See vom Baur [14].

In 1942 the Supreme Court rejected the government's long-drawn-out suit against Bethlehem Steel Company for the recapture of World War I profits it judged to be excessive. The Company made $24 million on $109 million of incentive contracts for ship construction—a margin over cost of 22 percent.

This triggered the adoption of renegotiation, incorporated into the Sixth Supplemental National Defense Appropriation Act,[8] and led to the prohibition of CPPC contracts. The new approach—renegotiation—authorized the procurement agencies to review each contract at settlement and to "re-price" those yielding "excessive" profits for recapture. In practice, profits were measured against cost although no base was established in the legislation. The Revenue Act of 1942 and the Renegotiation Act of 1943[9] (precursor of the present Act of 1951) shifted renegotiation from a contract-by-contract basis to a fiscal year basis, covering total profits from all contracts of a defense contractor. With no specific profit base in the Act the renegotiators in practice measured profits against cost and recaptured some $11 billion in excessive profits. The Act expired at the end of 1945.

Concern over excessive profits continued after the close of the war. Many in Congress were troubled by the practice of negotiating contracts and contract fees; the military services strongly supported the practice as the most feasible means of procuring the complex hardware spawned by changing technology. A compromise emerged in the form of the Armed Services Procurement Act of 1947.[10] It provided that purchases and contracts be made by formal advertising, except under certain circumstances. Then followed some 17 exceptions authorizing negotiated contracts. These embrace most procurements by dollar volume.

The Act imposed statutory limits on fees for CPFF contracts, the highest being 15 percent of estimated cost for experimental, developmental, and research, and 10 percent for most of the other CPFF contracts. The Department of Defense (DOD) established administrative limits on fees for all other types of negotiated contracts, as spelled out in the Armed Services Procurement Regulations (ASPR).[11] Before 1963 these fees generally ranged between 8 and 10 percent of estimated cost.[12] This high degree of uniformity came to be viewed by DOD as lacking in specific incentives to cut costs and reward performance. Accordingly, in 1963 DOD incorporated into ASPR the Weighted Guidelines to provide a greater percentage range of fees as an incentive to reduce cost and increase efficiency and performance. The range is 6 to 15 percent of estimated contract cost.

---

[7] Ch. 95, 48 Stat. 503 (1934).

[8] Ch. 247, Sec. 403, 55 Stat. 245 (1942).

[9] Ch. 619, Sec. 801, 56 Stat. 982; and Ch. 63, Sec. 701, 58 Stat. 78, respectively.

[10] 10 U.S.C. 2301 *et seq.*

[11] NASA has its own regulations, as does AEC, but like ASPR their fees are largely cost determined. ASPR is published in 32 Code of Federal Regulations.

[12] Fixed-price contracts based on competitive prices involve no cost analysis and no fee determination; whatever profits are made are neither reviewed nor controlled by the procurement agencies.

In Section 3-808.1, ASPR affirms DOD's policy of utilizing profit to stimulate efficient contract performance. The negotiation of very low profits (fees) or the use of historical averages or predetermined percentages provides no motivation to accomplish this purpose. Moreover, the best industrial capabilities are needed in defense work—low profit opportunities repel such capabilities.

To effect this general policy purpose negotiators are enjoined to establish a "profit objective" which would:

(1) Reward contractors undertaking more difficult work requiring sophisticated skills;
(2) Relate profit to the cost risk assumed;
(3) Reward contractors with an excellent past performance record and penalize those with a poor one; and
(4) Reward contractors who provide their own facilities and financing or whose competence is established through prior development work at their own risk.

The profit objective relies on detailed analysis of estimated contract cost, by categories of cost. In many cases past costs are a guide; on new and technologically complex products or systems these are of little help. The well-publicized cost overruns highlight this difficulty. Lacking is the relative stability and knowledge of cost characteristic of the regulated industries. When applicable, the Weighted Guidelines are used in the formulation of the profit objective. The weights are shown in Table 1.

When each of the cost components of Contractors Input [Item (1) in Table 1] is assigned a profit rate within each range and a dollar figure, the dollars of profit are totaled to derive the percentage of total cost.

> To this composite percentage, he [the negotiator] shall then add the specific percentages, assigned for cost risk, performance, and the other selected factors, to arrive at a total profit percentage. He shall then multiply the total recognized contract costs by this total profit percentage to determine the profit objective [ASPR 3-808.4 (e)].

This is a "cardinal principle" of the weighted guidelines method.

The Contractor's Input, or total cost, is of course crucial. In selecting a profit percentage for the Direct Materials cost component, these considerations are important: (1) the managerial and technical effort necessary to acquire the parts, subcontracted items, other materials, and special tooling, (2) the extent to which the contractor provides detailed specifications and creative design and managerial and technical help to his subcontractors, (3) the number of subcontracts to be let and monitored, and (4) whether new or established subcontractors are involved. In short, the issue is whether this aspect of the work is routine or highly complex.

The weights given to the Engineering and Manufacturing Labor components depend on the degree of specialized engineering and scientific talent required to fulfill the contract and on the manufactur-

TABLE 1

ASPR WEIGHT RANGES FOR PROFIT FACTORS

PROFIT FACTORS	PERCENTAGE WEIGHT RANGES
(1) CONTRACTOR'S INPUT TO TOTAL PERFORMANCE	
DIRECT MATERIALS	
PURCHASED PARTS	1 TO 4
SUBCONTRACTED ITEMS	1 TO 5
OTHER MATERIALS	1 TO 4
ENGINEERING LABOR	9 TO 15
ENGINEERING OVERHEAD	6 TO 9
MANUFACTURING LABOR	5 TO 9
MANUFACTURING OVERHEAD	4 TO 7
GENERAL AND ADMINISTRATIVE EXPENSE	6 TO 8
(2) CONTRACTOR'S ASSUMPTION OF CONTRACT COST RISK	0 TO 7
TYPE OF CONTRACT, REASONABLENESS OF COST ESTIMATES	
DIFFICULTY OF CONTRACT TASK	
(3) RECORD OF CONTRACTOR'S PERFORMANCE	–2 TO 2
SMALL BUSINESS PARTICIPATION	
MANAGEMENT, COST EFFICIENCY, COST RELIABILITY	
VALUE ENGINEERING ACCOMPLISHMENTS	
TIMELY DELIVERIES, QUALITY OF PRODUCT	
INVENTIVE AND DEVELOPMENTAL CONTRIBUTIONS	
LABOR SURPLUS AREA PARTICIPATION	
(4) SELECTED FACTORS	–2 TO 2
SOURCE OF RESOURCES	
GOVERNMENT OR CONTRACTOR	
SOURCE OF FINANCIAL AND MATERIAL RESOURCES	
SPECIAL ACHIEVEMENT	
OTHER	
(5) SPECIAL PROFIT CONSIDERATIONS	1 TO 4

ing skills and experience essential to the work. The various Overhead items vary in value depending on whether they are of a routine type or more specialized in nature.

As a separate profit component, the Contractor's Assumption of Contract Cost Risk [Item (2)] seeks to encourage contractors to shoulder risk by relating the profit ratio to the degree of risk assumed. The elements considered are type of contract, reliability of cost estimates, and the contractor's chances of success in performing under the contract.

> The extremes [in contract type] are a cost-plus-fixed-fee contract requiring only that the contractor use his best efforts to perform a task, and a [negotiated] firm price contract for a complex item.

The former entails minimum, or no, assumption of cost responsibility and therefore risk, the latter entails full assumption. Therefore,

> a zero rating shall be given to a proposed cost-plus-fixed-fee contract, and a 7 percent rating shall be given a closely priced firm fixed-price contract for a new, complex item [ASPR 8-808.5(c)].

The five other contract types fall within this range, each with its own narrow range.

Where, within these ranges, the negotiator will set the figure depends on his view of the reliability of the cost estimate and the contractor's past experience in costing and performance. Also the negotiator considers the relative ease or difficulty in performing the task under contract. Subcontracting may play a part in risk evaluation: subcontracting may add to a contractor's risk; or the contractor may drive a hard deal and shift risk from himself to the subcontractors.

The Record of Contractor's Performance [Item (3)] entails consideration of the company's "social" responsibilities and its performance as a business organization in meeting customer requirements. This factor has a value between plus and minus 2 percent of total costs. Its purpose is to "motivate contractors to improve their performance by rewarding them for excellent past performance and penalizing them for poor performance." (It is odd that a contract would be given a firm with a record of poor performance.) If performance is only "satisfactory" the contractor emerges with a zero in this evaluative category.

As for social responsibilities, contractors who strongly support the government's Small Business and Labor Surplus Area programs through subcontracting and technical assistance and job training for the hardcore unemployed are given favorable consideration in this area; those that do not supposedly earn a negative consideration. It might be noted that these social considerations generally play a minor part in the evaluation.

The important considerations relate to management, cost efficiency, reliability of cost estimates and delivery schedules, quality of product, and innovative and developmental contributions. There is considerable documentation in the procurement agencies on these elements and vast experience on the part of procurement officials with contractors and their performances. Out of this comes an informal rating system from excellent (despite an occasional flub), to quite good, to good, to fair. This evaluation is not easy; official views often differ; performance reports often are not informative. Somehow, out of this process of review, discussion, and bargaining with contractors, comes the evaluation—rarely negative, sometimes satisfactory, but mainly plus one.

The Selected Factors valuation [Item (4)] is intended

> (i) to discourage reliance upon Government resources, and (ii) to encourage contractors to outstanding performance. Therefore, in assigning a composite profit percentage within the range −2 to +2 for these Selected Factors, the Source of Resources factor will always be rated from 0 to −2 percent and the Special Achievement factor will always be rated from 0 to +2 percent. The Other factor may be rated with either a plus or a minus [ASPR 3-303.5(e)].

In this evaluation, government resources include only fixed capital, not progress payments.

The Weighted Guidelines thus set forth a range of percentages on specific components of cost to "value" these components, i.e., to derive the allowable profit on each cost component. These profits are summed to derive a composite percentage profit on total cost and the dollars of profit based specifically on this cost. Varying percentages are then applied to total cost to "value" the cost risk assumed by contractors, their record of past performance, their contribution of capital resources to the contract work, special achievement, and special profit considerations, e.g., rewarding contractors who perform developmental work before signing a contract to manufacture. This combination of "valuations" gives the composite percentage to be applied to total estimated contract cost to obtain the fee (profit).[13]

This, of course, is an *ex ante* estimate; the *ex post* contract audit tells the tale. There are questions of reasonableness of individual costs, the proper allocation of costs among a contractor's several or many contracts and between these and his commercial business, the limitations on, and disallowances of, costs, and the acceptability of the accounting system and procedures. In the end, the actual profit is generally less than the target or *ex ante* fee. Cost disallowances partly account for this (e.g., interest is not an allowable cost); and cost increases reduce the profit percentage. Cost estimation, cost behavior, and cost settlement are major problems for the procurement agencies, certainly in the procurement of new and technologically complex products and weapons systems that push the state of the art. Even if not so advanced, major items may take some years to construct with design changes introduced all along the line. To be sure, there are relatively standard items—weapons, ammunition, vehicles—that pose no great problems of cost. Often these items are purchased under fixed-price contracts, based on competitive prices, as is the whole range of housekeeping items. The costs and profits under these contracts are not the concern of procurement agencies, but they are likely to be reviewed in the renegotiation process.

In general, it may be said that costs, as a basis for fee determination in the defense-space industries, range from the reasonably predictable to the virtually unpredictable. The unknowns in a complex project are hazardous enough for the cost estimator, but the "unknown-unknowns" shatter estimates. The regulated industries have problems of their own in deriving cost estimates for rate determination by type of service. But there is a high degree of predictability as to type of cost and probable change in the near future. There are no unknowns of the magnitude that characterize much of government defense-space procurements. As a consequence, the rates established by cost-plus regulation yield profit outcomes that vary rather narrowly—when rate-setting authorities are responsive to change.

☐ **Some comments.** ASPR does not lay down a profit limitation policy in the sense of controlling profit outcome in accordance with a norm or norms. It sets forth a negotiated price policy, based on cost plus a controlled profit margin (fee) as a percentage of cost. The fee or margin is a component of price. In ASPR's terms a high or low profit means only a higher or lower profit margin over cost. At any given margin the dollars of profit vary with cost—the higher the cost the more dollars of profit, and vice versa. Given its cost-based fee policy, ASPR has no concept of profitability, no norm or norms except cost to guide its profit decisions. In this circumstance cost reduction by labor-saving investment is at best weak, at worst, absent. To be sure, there are competitive constraints in varying degrees but, for major procurements, probable contractor performance outweighs price in the final decision. Moreover, the procurement agencies have an interest in maintaining a reasonably diverse industrial base and, if possible, avoiding dependence on sole source suppliers. This reinforces the deterrent to cost reduction inherent in a cost-based fee policy.[14]

---

[13] The Guidelines seem not to have had the full effect intended. Many contractors assert that government negotiators often seem to ignore them and establish only the accustomed 8- to 10-percent fee. The negotiators are highly price conscious; many are reluctant to move to a higher percentage from a lower one, even when the Guidelines permit. The Guideline ranges give ample scope and "justification" for setting a composite percentage suitable to the negotiator. See Industry Advisory Council (IAC) Report to DOD [3].

Margins over cost per se reveal nothing about the profitability of a contract or the total contracts of a firm. Since profitability is usually measured as a ratio to capital employed, capital turnover is the key to this measurement. The profit margin multiplied by the turnover rate gives the measure of profitability. As will be seen, capital turnover varies considerably among contractors, yet ASPR is silent on this point. As a consequence its fee policy produces indeterminate profit outcomes.

Recent studies of defense-space contract profits and contract financing provide a wealth of data and conclusions in support of the above.[15] The General Accounting Office (GAO) Report shows that capital turnover rates on government work in its sample of firms averaged more than twice the turnover rate on these contractors' commercial work. For another sample of firms grouped in quartiles, the Industry Advisory Council (IAC) showed that turnover rates on government contract work ranged from 5.2 down to 2.2. A group of commercial firms ranged from 2.4 to 1.8. Averaging both groups, turnover rate on defense work was 3.2, on commercial work 2.0. Large contractors had higher capital turnover rates than smaller contractors. In summing up, the IAC Report stated:

> The profit determined via the present cost based method of calculation bears no discernible relationship to the competitive return on capital; and, since DOD profit policy focuses only on Profit/Sales while largely ignoring capital turnover Sales/Capital the result appears to be an inequitable distribution of profit among defense contractors.[16]

The higher turnover rates on defense-space business, compared with commercial work, is accompanied by lower average profit margins; generally profit margins on defense work average less than half the profit margin on commercial work. The lower margin on defense work is, in effect, offset by government-supplied capital, both progress payments and facilities. This offset in turn largely accounts for the higher turnover rates on contractor-owned capital.

The IAC Report is particularly concerned with the profit outcome disparities or inequities, measured against contractor investment, that abound in defense work. These disparities are largely rooted in the uneven financial support, via progress payments and facilities, given to defense contractors, leading to capital turnover differences.

In fiscal year 1971 total work-in-progress (inventory) on defense work came to about $9.4 billion; DOD investment in this inventory approximated $8.0 billion. But among contractors progress payments varied; large contractors received prompt payment upon billing the government, small contractors generally had to wait longer. Prime contractors on the average were paid in seven days after billing; in turn they paid off their subcontractors 33 days later, on the average. (By dollar volume, subcontracting amounted to 52 percent of the prime business.) On the average, subcontractors provide 22 percent of inventory financing, prime contractors only 5 percent, with the government providing most of the remaining inventory financing by way of progress payments.[17] Further inequities in profit outcome resulted from the uneven incidence of government-supplied physical facilities. Interest payments for borrowed funds are not allowable costs: these payments average 0.6 percent of total cost; small contractors and subcontractors bear relatively more of this burden than do the large contractors.

---

[14] Much, and perhaps too much, has been made of the Averch-Johnson (A-J) Effect in the regulated industries (see H. Averch and L. L. Johnson, "Behavior of the Firm Under Regulatory Constraint," *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–69). Here we have its opposite, the ID Effect of cost-based fee policy. Not to confuse the psychoanalysts, this is the Investment Deterrent Effect.

[15] See General Accounting Office (GAO) Report [11]. This is an *ex post* study of a sample of completed contracts. See also IAC Report [3].

[16] IAC Report [3], p. 40.

These profit disparities or inequities are inherent in the ASPR cost-based method of fee calculation, aggravated by progress payment practices. But a relatively consistent profit outcome is not the basic purpose of ASPR. Its objectives are improved performance and cost reduction through profit "incentives," i.e., variable margins on cost by cost categories. Attainment of this objective is most unlikely under a cost-based fee arrangement. In 1967 the Logistics Management Institute (LMI) in its report to DOD had this to say:

> Facility investments, soundly made, generally reduce total contract costs. Under present ASPR, facilities investment tends to lower rather than increase profit dollars on negotiated contracts. . . . Therefore, present Guidelines . . . tend to establish a lower dollar profit objective for an efficient plant . . . than it would for a less efficient plant producing the same output.[18]

In a similar vein the GAO Report states that

> . . . contractors in non-competitive situations are not provided with positive incentives to make investments in equipment that would increase efficiency and result in reduced cost, especially where follow-on contracts are involved. Under the present system . . . such investments tend to lower, rather than increase, profits in the long run.[19]

The IAC Report expresses the Investment Deterrent (ID) Effect more bluntly:

> . . . the current profit policy acts as a constraint on capital investment rather than as a stimulus. The present policy at best fails to adequately reward contractors for cost reductions, and, at worst, provides incentives for contractors to keep costs high and capital investment at the lowest possible level. . . . DOD profit policy, then, rewards most handsomely those contractors who maintain high labor-intensive operations while introducing only the minimum labor-saving capital additions required to obtain the business.[20]

It might be noted that, under ASPR's percentage differentials by cost categories, engineering labor is "valued" at 9 to 15 percent and manufacturing labor cost, only 5 to 9 percent. Labor intensity can thus be distorted; these differentials may account both for the strong demand for engineers when defense work booms and the sharp reduction in such employment when defense spending slacks off, as has happened since 1969. The differentials are dubious to begin with; the salary-wage differential established by the market may be presumed to roughly equate the "values" (productivity) on a per-dollar basis.

---

[17] *Ibid*; detailed data and IAC recommendations to remedy the inequities are given on pp. 1–42. An important recommendation is more uniform timing of progress payments. Another consideration: deferment of payments to primes until they meet their accounts payable to subs. Both recommendations were put into effect January 1, 1972. See *Wall Street Journal*, January 10, 1972, p. 4.

[18] LMI Profit Study [4], pp. 44–45.

[19] [11], p. 2.

[20] [3], p. 45.

The cost-determined fee basis for military procurement is the primary source of the perennial problem of cost control that has plagued procurement agencies since the advent of negotiated contracts. Although not immune to the cost-padding charge, the public utility industry has shown substantial rate reductions brought about by impressive cost-cutting capital investments over the same period of time. It is this contrast that plays a prominent part in the current moves to shift defense fee policy towards a partial capital base—a modified regulated-industry profit policy approach.

☐ **New directions.** In its 1967 report to the Department of Defense the LMI recommended a study of the feasibility of including capital employed as an element in fee determination. In 1969 the IAC (then Defense Industry Advisory Council) went on record favoring an adequate return on capital, together with cost, in fee determination. NASA contracted for a feasibility study of capital as a fee base.[21] The General Accounting Office Report recommended that "consideration should be given to capital requirements as well as such factors as risk, complexity of the work, and other management and performance factors." The IAC subcommittee report of June 1971 recommended the inclusion of contractor capital in the establishment of Department of Defense profit objectives. On July 17 Undersecretary David Packard approved the recommendation and requested an implementation plan by December 1, 1971. This plan culminates work in this direction begun by LMI in 1966.

Briefly, the IAC proposal covers these points:[22]

(1) Costs as developed in the Weighted Guidelines and capital are given equal weight in fee determination.
(2) Capital itself is shown as a weighted aggregate so that different degrees of risk in the capital components can be recognized and the form of capital with the most cost-reduction potential encouraged. The weights are (a) operating capital 0.7, (b) land 0.7, (c) buildings 1.0, and (d) equipment 2.0. The book value of each capital category is multiplied by the relevant factor to obtain "weighted capital employed."[23]
(3) Estimated contract cost divided by weighted capital employed gives the capital turnover rate.
(4) A predetermined rate of return on capital—the four-year average for a sample of durable goods companies in the FTC-SEC compilation—is the norm, with 3 percent added to adjust for unallowable costs. The FTC-SEC average is currently 22 percent, pre-tax. When the turnover rate is divided into the norm the calculation translates profit on capital to profits on cost.

---

[21] See Nash and Cibinic report [7]. The GAO Report [11] followed the Nash-Cibinic methodology in important respects.

[22] The details are given in the IAC Report [3], pp. 47–64.

[23] Weighting capital this way might be a special inducement to invest. But it analytically undercuts comparative profit analysis which is central to the IAC proposal, i.e., comparing yields on weighted capital with book capital yields is inadmissable.

(5) The Weighted Guidelines derivation of profit, and the profit on cost as derived just above are each weighted 0.5 to obtain the profit objective. The Guidelines are changed by the elimination of the category Source of Resources, which has proved "ineffective."

This proposal is a compromise between a completely capital-based and the present completely cost-based fee determination. The former is rejected as "undesirable" because a guaranteed capital return

> would provide no incentive to a contractor to increase his sales volume to fill costly excess capacity. A capital based profit policy is used in government regulated utilities where it is possible to relate capacity to a demand which can be forecasted with a fair amount of accuracy.[24]

The IAC Profit on Capital proposal is a landmark in defense procurement fee policy. While its primary aim is to cut procurement costs and lessen profit inequities, it is a halfway step towards articulating a profit policy. It seeks to distinguish several components of profit for that portion (one-half) which is capital related:

(1) An imputed interest, or "risk-free cost of money," applies to the book value of all capital components—working capital, land, buildings, and equipment.
(2) This rate, plus a risk premium, gives a "rental" charge for contractor buildings.
(3) The return on machinery and equipment is (a) imputed interest, (b) a risk premium, plus (c) an additional "reward" for inducing cost-saving investment, with the total (a,b,c) equal to the norm, i.e., the return on capital in the FTC-SEC sample of durable goods companies.

Only the excess rates over imputed interest [in (2) and (3)] represent "profit," a reward for risk-taking and cost-saving innovations.

For the portion of income covered by this proposal the policy analytically bears resemblance to regulated industry profit limitation policy: (1) the income components are distinguished; (2) they compensate certain functions (capital supply, the bearing of risk, cost-paring investment including possible innovations) performed by contractors; and (3) an objective norm guides and limits profit policy. By including innovative rewards, it is more liberal than regulated industry policy.

However, the cost base for profit determination, via the Guidelines, has a weight equal to the Profit on Capital procedure. The profit outcome as a whole will still be indeterminate, but to a lesser extent, and the shortcomings of the cost-based approach will remain, again to a lesser extent. The crucial question is whether the Profit on Capital approach will offset the investment disincentives inherent in the cost-based method. Which half will pull the stronger oar?

Another question arises: Is it appropriate to assign equal weights to cost and to capital to determine fees? Contractors range from relatively capital-intensive firms to predominately labor-intensive firms. To be sure, the new policy is designed to shift firms towards increased capital intensity. But there are severe constraints facing those performing only "service" under their contracts. The new formula may penalize the presently capital-intensive firms by assigning only a 50-percent weight to their capital in calculating the fee. The IAC recognizes the need for flexibility and experimentation. The variability of contractor production functions invites flexibility in the assigned weights.

---

[24] [3], p. 54.

A final comment: The proposed Profit on Capital approach defines, in a limited way, profit components and what they compensate; the cost-based approach does not. As a consequence, this is only a partial profit policy. That portion of the fee related to cost needs definition, i.e., a concept of functional compensation and the articulation of a norm or norms, before it can be said that a profit limitation policy exists in the defense-space procurement market.

3. The renegotiation process

■ Profit renegotiation is premised on the belief that negotiated contract fees may result in excessive profits, to be recaptured by an after-the-fact review of profit outcomes. Procurement in World War II was both hasty and hectic, precluding careful costing of the items bought. To guard against contract mistakes and contractor cupidity, Congress adopted the renegotiation technique; similar considerations prompted the passage of the Renegotiation Act of 1951 during the Korean War.[25] Conceived as a temporary, war emergency measure, renegotiation now rests largely on other grounds for its survival.

Briefly, the rationale runs this way. The defense-space procurement market differs in important respects from commercial markets. Military and space systems are produced in a market which usually lacks the cost-price guideposts of commercial markets, with their known technology and cost. Novel and technologically complex items are produced under great cost uncertainty; cost estimates frequently miss target. Even when the products purchased and prices paid are similar to those in commercial markets, unit costs may drop unexpectedly on volume, leading to windfall profits. This is one of the many "mistakes" procurement officials admit to in the execution of some 28,000 contracts in excess of $100,000 annually. And contract-by-contract pricing, with the fee arrangement, may produce instances of "excessive" profits given cost uncertainty, government-supplied capital, and ignorance of contractor-capital inputs.[26]

In addition, a more general point may be made: The cost-based fee arrangement results in indeterminate profit outcomes because of the diversity of capital turnover rates which are unknown to the government negotiators. The normal competitive pressures which adjust profit margins to some range of capital yields are largely absent in the defense-space sector; the margins (fee as a percentage of cost) are dominated by custom and mainly range from 8 to 10 percent of cost, although the permissible range is from 6 to 15 percent.

---

[25] See [13]. With ten extensions and considerable amendment it remains present law. The Act is administered by a five-member Board, appointed by the President with the consent of the Senate. The headquarters Board has two regional boards to conduct the detailed renegotiation work, subject to review by the Board.

[26] See Hartwig (Chairman, Renegotiation Board) [2], pp. 27–29.

As a consequence, profit outcomes may be fair, good, on up to "excessive" by the standard norms of profitability, depending on capital turnover. The partial shift to a return-on-capital base in setting contract fees may in time reduce the indeterminacy characteristic of past policy.

☐ **Statutory provisions.** The Renegotiation Act lists the following agencies whose contractors are (with exceptions) subject to renegotiation [Sec. 103(b)]: departments of Defense, Army, Navy, and Air Force, the Maritime Administration, the Federal Maritime Board, the General Services Administration, NASA, AEC, and the Federal Aviation Agency. Subcontractors of prime contractors are also covered. In both categories, firms with annual sales under one million dollars to the above agencies (as a group) are exempt from renegotiation. Firms above this "floor" must file annual financial reports with the Board. Other exemptions include standard commercial articles and services (under special conditions), raw materials and agricultural commodities, contracts and subcontracts with common carriers and public utilities, contracts not related to defense, and most competitively bid construction contracts. These exemptions shield several billion dollars annually from renegotiation; they are in the "non-renegotiable" category.

Sales to the government over and above these exemptions are "renegotiable;" the costs and profits associated with them are similarly designated. Contractors' annual reports must show financial data for both renegotiable and non-renegotiable business. The latter is mostly commercial business which comprises on the whole about 80 to 85 percent of the combined total.

Under the Act, renegotiable profit "means the excess of the amount received or accrued (under contracts and subcontracts) over the costs paid or accrued" allocable to such contracts. "The term 'excessive profits' means the portion of the profits . . . which is determined in accordance with this title to be excessive."[27] The Renegotiation Board makes this determination and recaptures the excess through (1) agreement with the contractors, or (2) by order when contractors refuse agreement. Contractors may petition the Court of Claims (previously the Tax Court) for a *de novo* redetermination.

In making its determination of excessive profits the Board must give "favorable recognition" to efficiency "with particular regard to attainment of quantity and quality production, reduction of costs, and economy" in the use of resources. In addition, the following factors must be considered:

(1) Reasonableness of costs and profits, with particular regard to volume of production, normal earnings, and comparison of war and peacetime products;

(2) Net worth, with particular regard to the amount and source of public and private capital employed;

(3) Extent of risk assumed, including the risk incident to reasonable pricing policies;

(4) Nature and extent of contribution to the defense effort, including inventive and developmental contribution and cooperation with the Government and other contractors in supplying technical assistance;

---

[27] [13], Sec. 103(e).

> (5) Character of business, including source and nature of materials, complexity of manufacturing technique, character and extent of subcontracting, and rate of turn-over;
>
> (6) Such other factors the consideration of which the public interest and fair and equitable dealing may require. . . .

The above is simply a catalogue of elements to be considered with no statutory guidance given the Board. This is both a strength and a weakness. The diversity among contractors requires flexibility in administration and in the exercise of judgment. But the absence of guidelines causes controversy.

A final provision needs mention. The Act provides for a five-year loss carry-forward.[28] The Board applies a contractor loss in any given year to its profits of the succeeding year. If the loss exceeds second-year profits the excess carries forward to the next fiscal year on up to five years. In fiscal year 1970 some 23 percent of the contractors reported losses aggregating $461 million.[29] For them this is an important provision.

The policy objective of the Act is the removal of "excessive profits." Unlike the regulated industry statutes and court opinions, the Act establishes no norm or range of norms and therefore gives no definition of excessive profits. The Tax Court has provided little help in this matter. In the relatively few cases that involved the relevant measurement of profit and its comparison with profit norms, the court has taken different, indeed inconsistent, positions.

☐ **The process.** Some 4,400 contractors filed financial reports in fiscal year 1970, with total renegotiable sales of $48 billion. Nearly 3,400 contractors made aggregate profits of $2.0 billion on a sales volume of almost $39 billion. Over 1,000 contractors lost money totaling $461 million on $9.3 billion of sales. Briefly, these reports are examined, first in a preliminary fashion, to "clear" those whose profits are negative or modest. Familiar financial ratios—ratios of profits to sales, to net worth, and to capital—tell the story. These ratios are compared with prior year ratios and with those of other contractors in the same industry. A cost review discloses the relative importance of "outside" (materials purchased and subcontracting) work; cost, sales, and profit allocations are checked. About 86 percent of the 1970 reporting contractors were cleared quickly at this preliminary stage. The remaining 14 percent (690) were assigned for detailed examination to one of the regional boards; the renegotiation process starts at this point and takes on the average nearly two years. If past experience holds, between 10 and 20 percent of this group will get excess profit determinations.

The first step at the regional renegotiation board is a reasonably detailed accounting review to establish the facts of each case. Renegotiable and non-renegotiable sales, costs, and profits are examined to assure a proper segregation. Costs are checked for allowability; balance sheet data are examined. Many accounting adjustments occur in this process, some to the benefit of the contractor, others not. For example, the regional boards completed 805 cases in 1970; profits were adjusted in 415, with a net upward adjustment amounting to $191 million. This increased the profits initially reported by 27 percent;[30] part of this increase was found to be excessive.

---

[28] *Ibid.*, Sec. 103(m)(4).

[29] Renegotiation Board Annual Report [12], p. 11.

Once the accounting facts are established the renegotiator appraises a contractor's profit position in light of his prior years' profit ratios, with other contractors in the same industry, and when appropriate with aggregate industry data from public sources. Ratios of profit to sales, net worth, and capital are compared with the contractor's ratios in his non-renegotiable work.

Other elements in profit assessment need brief comment. The ratio of material costs and subcontracting to total cost is an index of "work done" by a contractor. A high ratio indicates that the contractor is mainly an assembler of parts and components. A low ratio is taken to mean that the contractor "does more," his value added is relatively high, and he contributes a sizable capital with attendant risks. This contractor fares better than the high-ratio contractor in the judgmental process.

Contractor risk varies by type of contract. Fixed-price contracts yield both better profits and more losses than any other type. For example, the 1,000 firms reporting losses in 1970 had 58 percent of their renegotiable sales under fixed-price contracts accounting for nearly 77 percent of total losses. The profitable firms had 57 percent of their sales under fixed-price contracts on which they made 69 percent of their total profits.[31] Renegotiation takes into account the contract mix and accords more favorable treatment to contractors who work on fixed-price contracts than those who prefer the various less risky, cost-reimbursable types. Risk relates also to the source of capital; contractors using their own facilities and working capital bear more risk and are given more favorable treatment than those who use government facilities and progress payments.

Work of high complexity is given more favorable treatment than a routine production effort. Efficiency is also considered. Comparative cost data and procurement agency reports throw light on this matter; by and large, only the extremes are readily discernible. Contribution to the defense or space programs is difficult to assess; in practice the Board interprets this factor as meaning special contributions of an innovative type.

From this array of facts and judgments, and considering the diversity of the contractor groups, the Board comes up with a determination: to clear the renegotiated contractor, or to find excessive profits. At the regional board level contractors are heard; they may present additional facts and arguments. They may ask the headquarters Board to review a regional board finding of excessive profits and be heard again. All large cases are reviewed by the Board whether a review is requested or not. When the final decision is reached on the amount of excessive profit, the contractors may agree and pay, or not agree and receive a Board order to pay. At this point they may petition for a judicial redetermination.

Over the past 20 years, the Board has found excessive profits in over 4,000 cases; this is no more than 5 percent of the contractors who filed renegotiation reports with it. About 88 percent agreed to the finding and paid; some 480 did not and received a refund order. Nearly two-thirds of these then paid, the remainder petitioning the Tax Court. The court's redeterminations average about 5 percent below the original Board determinations. Through fiscal 1971 Board determinations of excessive profits totaled nearly $1.1 billion.[32]

---

[30] See Renegotiation Board Statement to the Government Activities Subcommittee [10].

[31] Renegotiation Board Annual Report [12], p. 11.

Renegotiation is a controversial process. Most contractors are loathe to part with profits deemed excessive by the Board despite the fact that they sign agreements to pay the refunds and only a small fraction have petitioned the Tax Court.[33] As the Act is drawn there is abundant room for honest differences of opinion as to the relative importance of the many elements to be considered by the Board. In the end a decision is a matter of judgment, made in a context that lacks statutory definitions, defined guidelines, profit norms, and consistent court precedent. With basic analytical issues unresolved, controversy abounds. This leads us to a brief review of these issues as revealed in Tax Court proceedings. These are the only "public" proceedings in the renegotiation process.

4. Some analytical issues

■ Since the Act establishes no base against which to measure profits and no norms against which to compare profit ratios, it offers no definition, however vague, of excessive profits. Among the statutory factors the Board must consider are net worth, capital employed and its sources, normal earnings, and reasonableness of costs and profits. These factors single out three possible bases against which profits may be measured: net worth, capital employed, and costs. Normal earnings may suggest a norm; but are these related to net worth, capital employed, or costs? Does the term "normal earnings" have much meaning in a context frequently devoid of "normality," or are normal earnings the average for industries or groups of firms in the normal business mainstream? These questions plague the renegotiation process because the Act provides no policy guideline or analytical framework. They surface to public view only in the proceedings of the Tax Court.

Three major cases arising from Korean War aircraft procurement but settled by the court in the 1960s illustrate the issues: Boeing Aircraft Company, North American Aviation, and LTV Aerospace Corporation.[34] See Table 2.

The Board found excessive profits in these instances; the companies petitioned the Tax Court for a redetermination. In its finding

---

[32] Total costs of the Board over these 20 years come to about $60 million; costs to business firms of handling renegotiation matters are unknown.

[33] There have been instances of contractors who offer a refund payment while being renegotiated; they knew their profits were excessive before a Board determination was made. Other contractors make voluntary refunds or price reductions during the year and before renegotiation. A sizable part of the $1.3 billion of such voluntary refunds may be attributed to a belief that their profits may be found excessive—and they prefer to avoid such a finding.

[34] Boeing Co. v. Renegotiation Board, 37 T.C. 613 (1962); North American Aviation, Inc. v. Renegotiation Board, 39 T.C. 207 (1962); LTV Aerospace Corp. v. Renegotiation Board, 51 T.C. (1968). The Board is the respondent in such cases; the Department of Justice handles the litigation, assembling evidence and witnesses. Board personnel are not involved because the proceedings are *de novo* and the redetermination is made by the court. See Burns [1].

TABLE 2

PROFIT RATES OF AIRCRAFT COMPANIES

COMPANY	RENEGOTIABLE SALES (MILLIONS)	RENEGOTIABLE PROFITS (PRE–TAX) (MILLIONS)	PROFITS AS PERCENTAGE OF	
			NET WORTH	SALES
BOEING (1952)	$700	$57	98	8.0
NAA (AV. 1953-1954)	640	50	81	7.7
LTV (1952)	45	5	176	10.0
LTV (1953)	70	9	243	13.0

of excessive profits the Board placed considerable weight on the high net worth returns; the companies relied on their profits-to-sales ratios as the basis for protesting the Board action. The appropriate base for the measurement of profit therefore figured prominently in these cases.

In the proceedings before the Tax Court the government increased its determination of excessive profits, at this point resting more heavily on the excessiveness of the net worth (capital) returns of these companies.[35] These returns were compared with samples of industrial firms (50 in number) taken to be representative of large-scale industry. In the years in question the pre-tax capital yield of these firms ranged from 22 to 24 percent. In this comparative profit analysis the government did not represent this average return as a norm to be applied to the three companies. The dispersion ranged up to a 40-percent return in the sample. Clearly an "excessive" profit could not be interpreted as any excess per se over the average.

In the court proceedings Boeing and NAA would have been left with capital yields ranging from 50 to 60 percent under the government's contention—from two to nearly three times the average of the sample firms. In terms of the average, these yields are high; in the end the distinction between a high profit and an "excessive" profit is a matter of judgment. LTV is a case in point; its capital yield would have exceeded 100 percent even with the government's upward revision of excessive profits. The yield is not the decisive element as it is in the regulated industries; the Act lays down other factors for consideration in a decision.

In the court proceedings both Boeing and North American argued the unreliability, inaccuracy, and unfairness of book net worth as a profit measurement base. They stayed with the profits-to-sales ratio; since this approximated the manufacturing industry average, they felt that their profits were reasonable. LTV argued further that neither the Act nor the court had established ceilings on net worth return or the sales margin and that as a small and growing firm its returns should not be compared with those of a group of large firms.

The court adopted one position in the Boeing and North American cases and an entirely different one in the LTV case.

In the former cases the court accepted the government's contention that net worth was the appropriate base against which to measure profits. Against the companies, it rejected sales as the appropriate base largely because they made extensive use of government-owned facilities. In doing so they minimized risk; in effect, they functioned as managers of government-owned capital. Thus, their profits-to-sales ratio cannot be compared to that of manufacturers who work with their own capital and bear the attendant risk. This disposed of the companies' profit measurement contention. At the same time the court rejected book net worth in favor of "adjusted net worth" as the true representation. The adjustment was the addition of the value of the companies' intangible assets—engineering design and know-how—to the value of book net worth. The court judged these to be equal to the value of the book assets. In effect it doubled net worth and halved the yield. It then found $13 million of excessive profits in Boeing, leaving it with 37 percent yield on adjusted net worth (77 percent on book net worth). Its findings of excessive profits in North American left the company with 35 percent on adjusted, and 68 percent on book, net worth. The court concluded that the 37 and 35 percent returns on adjusted net worth were reasonable in light of the returns of other similar airframe companies.[36]

---

[35] Book net worth was taken as the measure of company capital employed because of the absence of long-term debt.

In the LTV case, the court went along with essentially only one position taken in the other two cases, namely, that the profit-to-sales margin was an inappropriate measure for comparative purposes. It rejected book net worth as misleading and not helpful, and any adjustment of it as unwarranted for lack of evidence of the value of "know-how." It pushed aside comparisons of net worth returns—the government had offered no evidence that LTV was "like" the companies in the industry sample. In effect, the court in this case shattered the profit measurement ratios of both parties to the litigation. It left the measurement issue up in the air and comparative profit analysis in suspense.

More importantly, however, the court moved in a new direction. It distinguished two profit components: (1) the rental income of assets, and (2) an "entrepreneurial" profit. Implicitly it recognized a third: a payment for bearing risk. It said "a contractor which uses its own assets is entitled to an element of return equivalent to the rental value of the assets, a contractor using the assets of others is not." The use of government assets is a managerial function; a contractor is not entitled to a "rental" income from such assets, nor to a risk premium.

Entrepreneurial income is payment for the "productive" use of assets. Know-how, not separately measured, contributes to the productive use of assets, and the company is entitled "to an entrepreneurial profit for such contributions and not merely to a return on its own invested capital." A contractor is entitled to the same (rate of?) entrepreneurial profit on both its own capital and the government capital it uses. On its own capital it is entitled to a rental income, a risk premium, and an entrepreneurial profit; on governmental capital, only an entrepreneurial profit.

Basically, it is this functional analysis of profit which led the court in this case to reject both net worth (capital) yield and the sales margin as reliable measures of profit. The court opinion, how-

---

[36] This acceptance of comparative profit analysis is marred by the fact that adjusted net worth returns cannot properly be compared with book net worth returns.

P-5025

ever, did not follow up its own lead and offer alternative measurements. It did not apply its own promising analysis to a redetermination of LTV's excessive profits; it accepted the Board's prior determination.

The only clear point emerging from these juridical proceedings is the rejection of the profit-on-sales margin as the appropriate measure of profitability and the criterion for the assessment of profit. These opinions relate only to the cases under review; they may or may not set a precedent for subsequent decisions. The Board forbids the use of the profit margin as the criterion in judging whether or not profits are excessive, although it uses this ratio to signal cases that need careful review.

Adjusted net worth, in the sense developed in the Boeing case, has no place in the renegotiation process. Comparative profit analysis is essential; net worth adjustment destroys it. The court's rejection of it in LTV was salutory; but it broadened this rejection to include any net worth ratio and comparison. However, its "rental value" concept retains the capital base in determining one component of profit and perhaps also the risk premium. The implicit opportunity cost concept underlying a profit component on a contractor's own capital entails comparisons, but only those limited to the relevant portion of profit.

The entrepreneurial profit concept is vague; it is payment for the productive use of assets regardless of ownership. Whether the base for its measurement is total capital (contractor, government, or leased capital) or total sales is not clear. But as a component of profit it is presumably compensation for the management of capital, the efficiency of this effort, and perhaps any innovative developments that occur. Indeed, entrepreneurial profit in this sense could embrace all of the non-capital factors in the Renegotiation Act's list. Beyond this, the distinction between "rental" value and "entrepreneurial" profit as components bears some similarity to the impending change in DOD fee policy. The capital-based portion of the fee has its analogue in the "rental value-risk" profit component in the LTV case.

In summary, the Renegotiation Act and the Tax Court decisions offer no analytical framework to guide the Renegotiation Board. As a consequence, the administration of the renegotiation process is a difficult and controversial task. Even the elemental matter of profit measurement is subject to dispute, as the Tax Court proceedings vividly demonstrate. And the diversity of ill-defined statutory factors listed in the Act invites dispute on the Board's interpretation of them. In contrast to the profit limitation policy in the regulated industries, the renegotiation process at best works with an inchoate policy. It needs a reasonably clear answer to the basic question underlying negotiated profit (indeed, all government-controlled profit): What are these profits paid for? What functions or services are compensated by the fees, and the resultant profits, in the military-space procurement market?

5. Concluding remarks

■ The profit-fee policy in the government procurement market is in trouble and has been since this market emerged as a major element in the U. S. economy. As policy it is disjointed; it lacks clear-cut profit guidelines. At the procurement phase, contractors often complain about inadequate profits. Some critics take the other tack, charging contractors with making exorbitant profits and the procurement agencies with laxity. Renegotiation has its share of critics, some complaining that the Board recaptures too much profit, others, not enough. There is a large grain of truth in some of these allegations. The IAC Report admits to the inequities and inefficiency of cost-based fee policy; this is a rare display of candor coming after more than 30 years and the expenditure of more than a trillion dollars under this contract fee policy. No similar self-evaluation is under way at the Board, but one is under consideration.

The partial shift of DOD policy towards relating negotiated fees to a capital base should lessen the indeterminacy of profit outcomes inherent in its cost-based approach. And it might mitigate the Investment Deterrent Effect. Only time will tell; there are conflicting pulls in the combined policy, and *ex ante* capital estimation for complex projects might prove unworkable. The remaining cost-based component of the negotiated fee perpetuates in half-measure the disabilities of past fee policy. Only the *ex post* review will bring the profit outcome under a profit limitation policy.

As for renegotiation, its inchoate profit policy needs analytical sharpening to distinguish the functions or services compensated by the profit allowed. Indeed, the LTV opinion of the Tax Court points the way, as does the new DOD policy. The Renegotiation Act lists the elements to be considered in the Board's decisions; they need the (flexible) embrace of an analytical framework. These elements fall into the following general categories: (a) the supply of capital, (b) the bearing of risk, (c) the organizational-managerial function of the corporate entity per se, and (d) innovation, in product or process. The following approach is suggested:

(1) An imputed interest on total book capital, the "rental value" of its physical assets, and graduated risk premiums, cover (a) and (b). The allowable rates can be approximated from capital market and leasing company information.

(2) In addition to its own capital, a contractor normally employs some combination of (a) leased assets, (b) government facilities, and (c) fully depreciated but useful capital. The value of these assets needs estimation, a difficult but not insuperable problem. To the value of these assets, together with the contractor's book value of plant and equipment (total capital employed), allow a rate of return (perhaps 2 or 3 percent) to compensate the organizational-managerial function performed by the corporate entity. This is analagous to the "entrepreneurial profit" set forth in the LTV opinion.

The sum of the dollars derived from (1) and (2) above would then constitute a norm to guide Board decisions. When actual profits equal or fall short of this norm, the contractor is cleared. When profits exceed this norm, they need careful analysis. If the facts at hand disclose special innovative development or special efficiency achievements (neither is common), all or some portion of the excess could be allowed in the Board's judgment. If the facts disclose evidence of gross inefficiency (also not common), the amount allowed could be less than the norm. All of this entails judgment, of course, but renegotiation is a judgmental process as it is. The suggested approach provides presently missing parameters for the exercise of judgment.

The regulated industries long since have reached the stage of profit limitation policy hedged in by norms. Profit is conceived mainly as capital yield and a premium for risk, both measured against a capital base. The norm for allowable profit is, in general, given by the capital market, with the allowable yield rate sufficient to attract capital. It is a complex process and possibly too restrictive in some important respects, possibly needless and too costly in others, but in contrast to the government procurement market it is an articulated policy of profit limitation.

The present moves by DOD, and a shift to a set of norms for renegotiation, such as suggested above or a modification of it, would achieve a similar articulation but with more variables to account for (1) a greater range of functions and, therefore, profit components, and (2) the heterogeneity of the defense-space sector compared with the regulated industries.

Whatever its merits may be (and this is subject to question) profit control needs a rationale and reasonably quantitative guidelines. Otherwise it is capricious. Why is this much profit allowed, rather than more or less? A delineation of the functions or services to be compensated by components of profit provides guidelines and an answer.

How does this relate, if at all, to the current interest in incomes policy? In bits and pieces government has long intervened with efforts at income control in the form of floors and ceilings (roofs), at times in combination with entry barriers to form shelters. Examples include segmented profits controls, social insurance and welfare, minimum wage legislation, direct subsidies, and price support for agriculture. This formless structure of intervention may now be extended and given an architectural design to contain inflation. Apart from the fact that incomes policies rarely work well or for long in a context of contending political parties attentive to pressure groups, any additional control superimposed upon present profit control would fail the equity test. The source of trouble lies elsewhere.

References

1. BURNS, A. E. "The Tax Court and Profit Renegotiation." *The Journal of Law and Economics,* Vol. 13, No. 2 (October 1970), pp. 307–26.
2. HARTWIG, L. E. "The Need for Renegotiation." *Defense Management Journal,* Vol. 6, No. 2 (Spring 1970), pp. 27–29.
3. INDUSTRY ADVISORY COUNCIL. *Report to the Department of Defense on Defense Industry Contract Financing.* Mimeographed. Washington, D. C., June 1971.
4. LOGISTICS MANAGEMENT INSTITUTE. *Profit Study. Report to the Department of Defense.* Washington, D. C., 1967.
5. MACAVOY, P. W. "The Effectiveness of the Federal Power Commission." *The Bell Journal of Economics and Management Science,* Vol. 1, No. 2, (Autumn 1970), pp. 271–303.
6. ———. "The Formal Work-Product of the Federal Power Commissioners." *The Bell Journal of Economics and Management Science,* Vol. 2, No. 1 (Spring 1971), pp. 379–95.
7. NASH, R. C. and CIBINIC, J., JR. *Profit Planning.* Looseleaf. Washington, D. C.: George Washington University, 1968.

8. Posner, R. A. "Natural Monopoly and its Regulation." *Stanford Law Review*, Vol. 21, No. 3 (February 1969), pp. 548–643.
9. Stigler, G. J. and Friedland, C. "What Can Regulators Regulate? The Case of Electricity." *The Journal of Law and Economics*, Vol. 5, No. 2 (October 1962), pp. 1–16.
10. U. S. Congress, House, Committee on Government Operations. *Statement Submitted by the Renegotiation Board to the Government Activities Subcommittee.* May 14, 1971.
11. U. S. General Accounting Office. *Defense Industry Profit Study: Report to the Congress by the Comptroller General of the United States.* Washington, D. C.: U. S. Government Printing Office, March 1971.
12. U. S. Renegotiation Board. *Annual Report.* Washington, D. C.: U. S. Government Printing Office, 1970.
13. *U. S. Statutes at Large.* Renegotiation Act of 1951. Ch. 15, 65 Stat. 7 (codified in 50 U.S.C. App. Spec. 1211 *et seq.*).
14. vom Baur, F. T. "Fifty Years of Government Contract Law." *Federal Bar Journal*, Vol. 29, No. 4 (Fall 1970), pp. 305–58.

Cost allocation in railroad regulation

Zvi Griliches

Professor of Economics
Harvard University

This paper reviews a number of cost studies conducted by the Cost Finding Section of the ICC purporting to show that the elasticity of total railroad costs with respect to output (percent variable) is 0.8.

These studies are criticized primarily for using deflated and also irrelevant data. Deflation by miles of track leads them to evaluate this elasticity as the conceptually wrong "average" railroad level. In the course of the paper the statistical reasons for and against using deflated variables are explored in some detail and alternative computations are presented indicating that the same data are consistent with constant or only mildly increasing returns to scale to a proportionate expansion of traffic. Several published econometric studies of railroad costs are also reviewed and found to be consistent with the absence of significant increasing returns to scale to an indiscriminate expansion of traffic.

1. Introduction

■ In 1962 the ICC instituted proceedings on "Rules to Govern the Assembling and Presenting of Cost Evidence," which became known as Docket No. 34013.[1] One of the suggestions championed by the Cost Section of the ICC in these proceedings was that it be accepted that "on average" 80 percent of the railroad costs are "variable" (out-of-pocket). The usefulness of such a magic number was presumably in reducing the range of arguments before the ICC and in allowing the railroads to reduce some rates by 20 percent below "fully distributed costs" without having to justify it in detail and without it being adjudged *prima facie* as "unfair competition."

While the goal of more flexible pricing policies and less litigation may have been very desirable, the arguments in support of the 80-percent number were rather weak. They are outlined in the statement of H. J. Day,[2] and are based in part on materials summarized in an earlier ICC publication, *Explanation of Rail Cost Finding Procedures and Principles Relating to the Use of Costs.*[3] In essence, the Cost Section of the ICC had run a variety of linear regressions relating total costs per mile to tons per mile and claimed that the results of these regressions supported the 80-percent figure.

The paper that follows contains an appraisal of the evidence presented by the Cost Section and some related econometric studies.

Zvi Griliches received his Ph.D. in economics from the University of Chicago, where he also served on the faculty of the Department of Economics for 13 years. He is a Fellow of the American Academy of Arts and Sciences and was awarded the John Bates Clark Medal in 1966 by the American Economic Association. Professor Griliches is also a Fellow of the Econometric Society, a member of its Council, and coeditor of its journal. He has published numerous articles in the fields of econometrics and technological change.

[1] See [11].

[2] See [12].

[3] See [10].

This appraisal was included in Docket No. 34013 in somewhat different form as part of the submissions by the motor carriers. It illustrates both the usefulness and the difficulties associated with statistical and econometric methods of cost allocation. The relevant percentages of railroad costs that are variable depend critically on the context of the question, the choice of observations, and the statistical procedure used. At the average levels of traffic that appear to be most relevant the percent variable is above 95. These results agree with those from econometric studies.

On July 13, 1970, the ICC in Docket No. 34013 "found" that "approval and adoption of specific cost formulas, with a view toward giving prima facie validity to formula-based costs [were] not shown to be necessary or desirable." In addition, the ICC also supported our conclusion that regulation requires sophisticated and detailed analysis by calling for the establishment of a task force to consider further research into guidelines for cost allocation. (See the Appendix for a more complete extract from this decision.) It remains to be seen whether the simplistic methods of allocation rejected by the ICC will be replaced by methods of allocation that reflect the underlying complexity of relationships between cost and scale in railroad operations.

•

☐ **What is to be measured?** All of the studies to be examined attempt to establish whether or not the railroad industry is operating under conditions of increasing returns to scale or decreasing cost, and to measure the magnitude of these returns to scale.[4] To do this, they estimate a parameter (number) called "percent variable." Therefore, before we proceed further, we must examine what it is that these studies attempt to measure. Only having done this will we be in a position to evaluate how well they have succeeded.

The "percent variable" concept is traditionally defined as incremental cost divided by total cost. In the language of economic theory, the same concept can be expressed in per-unit-of-output terms as the ratio of marginal cost to average cost. It can also be interpreted as the elasticity of total cost with respect to a change in output, i.e., the percentage change in cost due to a one-percent change in output. Assuming that there exists a well-defined relationship between cost and output, we can write the total cost function as

$$c = f(x),$$

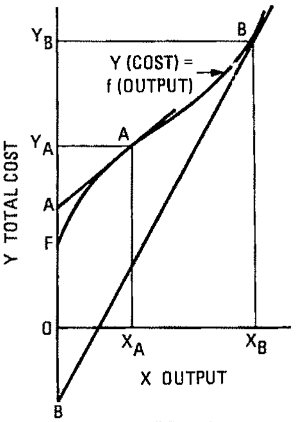
marginal cost as dc/dx, and average cost as c/x. In this notation

$$\text{percent variable} = \text{p.v.} = \frac{MC}{AC} = \frac{dc}{dx}\frac{x}{c}.$$

Now, conventional economic theory envisages a cost to output relationship something like the one depicted in Figure 1. There is a fixed level of cost, *OF*, which has to be incurred if one is to produce

FIGURE 1

A CURVILINEAR COST FUNCTION



$$\text{PERCENT VARIABLE}_A = \frac{\text{MARGINAL COST AT A}}{\text{AVERAGE COST AT A}} = \frac{AY_A}{0Y_A} < 1$$

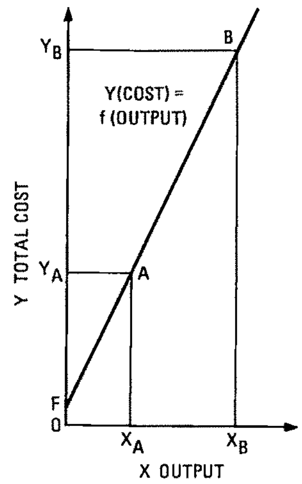
$$\text{PERCENT VARIABLE}_B = \frac{\text{MARGINAL COST AT B}}{\text{AVERAGE COST AT B}} = \frac{BY_B}{0Y_B} = \frac{0Y_B + 0B}{0Y_B} > 1$$

---

[4] The studies examined here are mainly those prepared by the Cost Section of the ICC. We shall treat in detail only the studies reported in the statement of H. J. Day [12], though most of our comments apply also to the studies reported in *Explanation of Rail Cost Finding Procedures* [10]. In addition, we shall also discuss briefly the several published econometric studies of Borts and Klein [2,3,4] and [6], respectively.

FIGURE 2

A LINEAR COST FUNCTION



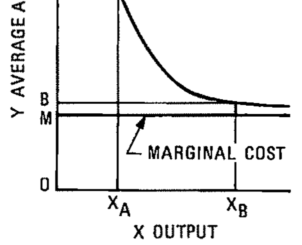
$$\text{PERCENT VARIABLE}_A = \frac{\text{MARGINAL COST AT A}}{\text{AVERAGE COST AT A}} = \frac{OY_A - OF}{OY_A}$$

$$\text{PERCENT VARIABLE}_B = \frac{\text{MARGINAL COST AT B}}{\text{AVERAGE COST AT B}} = \frac{OY_B - OF}{OY_B}$$

$$1 > PV_B > PV_A$$

FIGURE 3

AVERAGE AND MARGINAL COST CURVES (ASSUMING LINEAR TOTAL COST CURVE)



$$PV_A = \frac{OM}{OA}, \quad PV_B = \frac{OM}{OB}$$

any x and which is independent of the actual level of x. Costs rise with output, but only slowly at first. Eventually, bottlenecks of various sorts are reached and costs begin to rise more rapidly with output. Thus at x_A, percent variable is equal to AY_A/OY_A and is less than 100, while at x_B, it equals BY_B/OY_B and is larger than 100. For a curvilinear relationship both marginal cost and average cost changes with output, and so also does percent variable.

The Cost Section studies assume that the true relationship between output and total cost is linear rather than curvilinear. This is an assumption which, as far as I am aware, has not been tested by the Cost Section. Given this assumption, the cost to output relationship is like the one depicted in Figure 2. For such a linear cost function, marginal cost is constant, independent of output, and equal to the slope of the function. Average cost, however, is not constant. In fact, it falls continuously with output. Figure 3 illustrates this point, translating Figure 2 into per-unit-of-output units. In general, for linear cost functions, whether or not percent variable (p.v.) is smaller or larger than 100 depends on whether the intercept (OF) is positive or negative. In either case, percent variable is *variable*, it does not remain the same at different levels of output, and it approaches 100 as the output rate gets large.[5]

The main point of this digression is that for almost any type of cost function, linear or nonlinear, it is extremely important to decide on the level of output at which percent variable will be measured.[6] Since it is not the same along the curve, the answer will depend crucially at what point the question is asked, and it will be different for different situations. Thus, if one is interested in this number for some "average" level of output, it is crucial to specify what one means by this average, since the answer is extremely sensitive to the choice of a particular "average."

For example, the Cost Section estimates by regression methods that in 1958 the incremental (marginal) cost of a ton-mile was approximately 6.64 mills (thousandths of a dollar). At the same time it estimates that average cost was 8.56 mills, which leads to its estimate of 77.6 as the percent variable. But this "average" cost is computed by averaging the cost per ton-mile for each of about one hundred railroads, with about half of them being very small but accounting for half of the weight in this average. That is, the Cost Section takes the average per-ton-mile cost experience of the smallest railroad and the largest and gives it equal weight. Doing this, its "average" does not represent the experience of the "average" ton, or average ton-mile, or even of the average railroad, unless one considers one sparrow and one horse as comparable. Using the same data, we find that in 1958 total costs were about $89 million per railroad and that gross ton-miles came to about 13 billion tons per railroad. The "average" ton, in the sense that each ton gets equal weight in this average, moved at an approximate average cost of 6.64(88,886/13,381) mills per ton. Thus, at this "average," marginal cost is about equal to average, and the percent variable equals 100 percent. Actually, if one recomputed the marginal costs accordingly (again giving the small roads smaller weight), they too fall to 6.47, leading to a corrected estimate of 97.4 percent.

---

[5] See Borts [3] for a more detailed discussion of some of these points.

[6] Only for the cost function $C = AX^b$, which is linear in the logarithms of C and X, is the percent variable a constant at all levels of output. In this case, $dc/dx = bAX^{b-1}$, $c/x = AX^b/x$, and $(dc/dx)(x/c) = b$.

The same sort of confusion about what is the right "average" can be found in the Day statement[7] where a number of percent variable findings for the 1939–1946 period are summarized for different account groups. These range from 32 to 123 and a median value is reported as 79. But these various account groups account for very different fractions of total costs. In 1959, for example, the 32 category (traffic expenses) accounted for only about one percent of total operating expenses, while the 123 category (transportation, rail line) was responsible for half of the total.[8] Using such expense shares as weights, the "average" weighted percent variable from the same numbers is 100 and not 79.

To summarize this point, it is very difficult to proceed to a discussion of the correct measurement of "percent variable" unless it is possible to agree on *which* percent variable one is interested in—which "average," for what railroads, and at what traffic density. In what follows, I shall assume that the desired percent variable is defined for the average ton moved giving, in this sense, the larger railroads a larger weight in it. This does not mean that the statistical relationship must necessarily be computed using the same weighting scheme (that is a separate problem and will be discussed below), but whichever way the cost-output relationship is estimated, the percent variable is to be computed (derived from it) at the average ton level.

2. The cost-output relationship: problems of specification and measurement

■ For a particular relationship to be a good estimate of what it purports to measure, (1) it has to be defined correctly, (2) the variables have to be measured in relevant units, (3) it must include explicitly all of the important variables affecting the relationship, and (4) the excluded "minor" variables must be distributed randomly and must not be correlated with levels of the variables which are of particular interest to us. As far as I am aware, no serious investigation was made by the Cost Section as to whether any of these criteria are satisfied by its measurement procedure.

The number of problems is legion. First there is the problem of aggregation over different types of cost and output, e.g., freight and passenger, maintenance, and line transportation. For the aggregation not to lead to distortions, either the various subsets should exhibit the same cost to output relationship, which is contradicted by the evidence assembled in *Explanation of Rail Cost Finding Procedures*, or the share of these components should be the same for all railroads. On the latter point no evidence is presented. In general, the estimates based on such an aggregation can only be defended if the proposed increases in traffic will be distributed among railroads and among types of output in the same proportions as were the average outputs in the past.

CALCUTTA UNIVERSITY CENTRAL LIBRARY

Besides the question of whether or not such relationships should be computed separately for passenger and freight traffic, for line-haul and switching operations, and so forth, there is the very serious question of what is the relevant output unit. The Cost Section uses gross ton-mile per mile as the relevant unit. Cost may, however, be affected more by the number of freight cars moved rather than by the average weight of freight per car. It is doubtful whether the changes in level and characteristics of output can be adequately summarized by one measure. Thus, a more careful study[9] uses three separate variables—loaded freight car miles, carloads of freight, and empty freight car miles—as its measures of output in the line-haul part of railroad activity.

---

[7] See [12], p. 7.

[8] See *Statistical Abstract of the U. S* [9], Table No 755, p. 581.

What we are after is an estimate of a long-run relationship, in some sense, between cost and output. This is one reason for preferring cross-sectional estimates to time series estimates, since the latter are affected very strongly by short-run and presumably irrelevant fluctuations in costs and output.[10] But looking at a one-year cross section of data for a number of railroads does not assure us that all of the observations are long run. In any one year, part of the observed output and cost levels are "transitory," the result of short-run influences or lags in adjustment. That is, what we would want to observe are data on the permanent components of cost (C_p) and output (X_p). What we actually observe is subject to measurement error and to transitory influences. Hence we can write the observed variables as the sum of two components—the systematic or permanent part and the transitory or error part:

$$C = C_p + C_T, \quad X = X_p + X_T.$$

The relationship we are after presumably exists only between the "permanent" parts of these variables. Now, let us assume that the transitory parts are random, zero on the average, and uncorrelated with each other or with the systematic components. If we assume, in addition, that the true relationship is $C_p = kX_p$ (i.e., p.v. = 100), but that we estimate

$$C = a + bX$$

using the observed variables (which are subject to these random influences), we can write the estimated coefficients as

$$b = \frac{\text{covariance } CX}{\text{variance } X} = \frac{\text{cov} C_p X_p}{\text{var} X_p + \text{var} X_t} = \frac{k \text{ var} X_p}{\text{var} X_p + \text{var} X_t} = kP_x,$$

where

$$P_x = \frac{\text{var} X_p}{\text{var} X_p + \text{var} X_t} < 1 .$$

P_x is the fraction of total variance of observed output accounted for by the fluctuation in the systematic component. Thus, as long as there is any transitory variance in X, the estimated b will be smaller than the "true" k. Similarly, even though we assumed that there is no intercept in the relation connecting the true (permanent) variables, the estimated intercept

---

[9] See Borts [4].

[10] On the other hand, the use of cross sections raises the question of whether variables specific to particular railroads have been included in the relationship.

$$a = \bar{c} - b\bar{x} = \bar{C}_p - kP_x\bar{X}_p = k\bar{X}_p(1 - P_x)$$

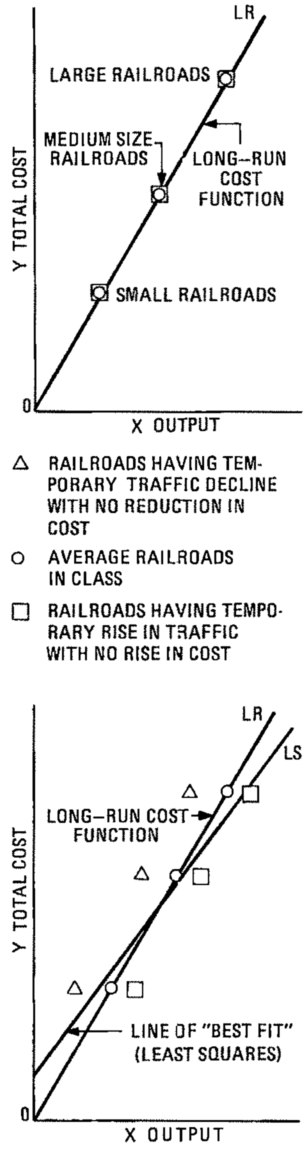
will be positive. Moreover, the estimated elasticity of costs, or percent variable, will equal

$$b\bar{x}/\bar{c} = kP_x\bar{X}_p/k\bar{X}_p = P_x,$$

or just the fraction of the systematic component in the total variance of output. In this example, the computed percent variable turns out not to have any relationship to the true cost-output function. This is the result of the particular assumptions; but it is true more generally that if our variables are subject to such short-run influences, the estimated elasticity (percent variable) will be biased downward. This argument is illustrated graphically in Figure 4.

FIGURE 4

INFLUENCE OF SHORT-RUN FLUCTUATIONS ON THE ESTIMATED COST FUNCTION



This bias can be reduced by using as our observations periods longer than one year. This is why I chose in reexamining the Cost Section data to recompute the relationship using the 1957-1961 five-year averages as the units of observation.[11]

The Cost Section "deflates" its observations by miles of road (M). Such a deflation by a "size" variable in a cross section can serve two purposes. To the extent that total costs are also a function of size and not just of total tons of traffic moved, this may allow one to reduce the problem from that of three variables (C,X,M) to that of two (C/M and X/M). Also, it reduces the influence of sheer size on the estimates. In the undeflated form they could be dominated by several extremely large observations, leading to spuriously high correlation coefficients. The cure, however, may be worse than the disease.

> The main point can be simply summarized: correlating ratio variables and making inferences from these . . . is extremely hazardous business. . . . If primary interest centers on the simple correlation between the numerator series, it would usually be appropriate to proceed in a straightforward manner by relating these values to one another.[12]

To use deflation to eliminate the size component, one must assume that miles of road are in fact the relevant size measure *and* that the cost relationship is homogeneous in output and size, i.e., that there are no costs which are independent of size. Implicitly, the Cost Section assumes that the true relationship is given by

$$C = aM + bX, \qquad (1)$$

which upon division by M becomes

$$C/M = a + b(X/M). \qquad (2)$$

But if the true relationship is

$$C = aM + bX + c, \qquad (3)$$

deflation leads to

$$C/M = a + b(X/M) + c/M. \qquad (4)$$

---

[11] Actually, this turns out to make very little difference and is a reflection of the fact that the separate year estimates are very much alike. I shall return to this point below.

[12] Kuh and Meyer [7], p. 403.

The assumptions underlying the Cost Section procedure imply that the coefficient a is "significant" while the coefficient c is insignificantly different from zero. These assumptions can be tested by fitting equations (3) and (4) to the same data instead of the equation (2) form chosen by the Section.

For the 97 railroads for which observations were given in all of the five years, we find, using 1957–1969 averages as our units, that

$$C/M = \underset{(6218)}{13{,}010} + \underset{(0.871)}{6.431}\, X/M \qquad R^2 = 0.365, \quad (5)$$

$$C/M = \underset{(5115)}{827} + \underset{(0.682)}{6.439}\, X/M + \underset{(393{,}000)}{3{,}065{,}000}\, 1/M \qquad R^2 = 0.614, \quad (6)$$

$$C = -\underset{(2.906)}{1.884}\, M + \underset{(0.375)}{6.613}\, X + \underset{(4730)}{3676} \qquad R^2 = 0.945, \quad (7)$$

with a significant c coefficient in equation (6) and insignificant a coefficients in equations (6) and (7). Thus, the assumptions underlying the deflation for size are contradicted. There is no evidence that M belongs in the equation in any form. It appears in a "significant" form in equation (6) only because the other variables were divided by it. This may be a reflection of the fact that the total miles of road are a poor measure of the "size" of the railroad. Anyway, division by an irrelevant variable is uncalled-for.

Another related argument for deflation is to minimize the influence of the extremely large observations on the estimated equation. Implicitly it is assumed that the larger the observation, the larger is the error associated with it, and that by dividing them by a size measure one gets numbers whose errors are roughly comparable to each other; that is, deflation is performed to stabilize the error variance. Note, however, that an unstable error variance does not by itself lead to biased coefficients. They are still consistent estimators of the "true" parameters. But stabilizing the variance should lead to more efficient estimators, estimators having a lower dispersion.

In technical language, it is assumed that the true relationship is

$$Y = a + bX + u,$$

where $u_t = e_t M_t$ is random error which is proportional to size (and e is a truly random independently distributed variable with zero mean and constant variance σ^2). Then $Eu_t^2 = \sigma^2 M_t^2$. In this case, the "best" (most efficient) estimates can be found by minimizing

$$\sum \frac{1}{\sigma_{u_t}^2}(Y - a - bX)^2 \quad \text{or} \quad \sum \frac{1}{M_t^2}(Y - a - bX)^2 .$$

This is almost, but not quite, the same as using Y/M and X/M as variables. The correct weighted least-squares procedure is equivalent to finding the a and b which will minimize

$$\sum \left(\frac{Y}{M} - \frac{a}{M} - b\frac{X}{M} \right)^2 ,$$

while estimates using the deflated data are derived by minimizing

$$\sum \left(\frac{Y}{M} - a - b\frac{Y}{M} \right)^2$$

and do not have the a/M term in them. This leads to a substantial difference in the estimates unless $a = 0$. The Cost Section method of deflating by M leads to an estimate of 75 as the percent variable, while weighting the absolute regression by $1/M^2$ or computing $Y/M = a/M + b(X/M)$ leads to 96.5 as the estimated percent variable.

While deflating as a proxy for weighting may be efficient in the case of size associated error variance, deflating by miles may be overdoing it. An examination of the scatter of cost on tons does not indicate that variance grows markedly with size. While one may not want to give the Santa Fe 29 times the weight of the Belt Railway of Chicago, this does not mean that one should treat them equally. Besides, there are other cures for this disease. Instead of deflating by size, one can perform one of several possible variance stabilizing transformations. For example, instead of using the absolute levels of the variables, one can use their logarithms. This will reduce the weight given to very large observations.

It is important to remember the reason for all this weighting or deflation: to improve the efficiency of the estimated regression line. Once this line has been estimated, it does not follow at all that the average elasticity (percent variable) should be computed at the average of the deflated or otherwise weighted observations. The choice of weights to minimize errors in the estimation procedure has nothing to do with the choice of the "average" (and associated weights) for which to compute the percent variable. While one may get better estimates using per-mile figures, the relevant density at which the percent variable should be computed is not at the unweighted average of the average densities for all of the railroads.

3. Some alternative computations

■ Because of limitations of time and data we shall beg most of the problems raised above and report only on some recomputations using the same Cost Section data and the same definition of output and cost.

As noted above, to reduce the possible influence of short-run variations we averaged the data for the available five-year period 1957–1961.[13] Also, an examination of scatter diagrams of the data indicated clearly that the very small roads had very much higher average costs, that they were distributed much more randomly (i.e., the error variance in the per-mile form was much larger for them), and that they appeared to be lying along a different cost function or along a distinctly different segment (in slope and shape) of the cost function. To take this into account, I divided the sample, arbitrarily, into two groups: roads which operated less than 500 miles of track in 1957, and roads which operated more than 500 miles. This divides the sample roughly into two halves having 45 and 52 observations each. Even though the small roads account for about half of the total observations, they account for only about 5 percent of total cost in the railroad industry. The second group of 52 roads accounts for 95 percent of the "volume" of the industry and is clearly the most relevant for our purposes.

---

[13] This left us with 97 roads in the sample. The fact that prices changed over this period does not matter, since (1) only the dependent variable is affected by this, and (2) the same price index would probably have been used for all the roads, leading to a small (constant) difference in the coefficients.

We concluded in the earlier section that "miles" is not a good size variable and does not belong as a variable in the cost-output equation. We also concluded that there may be an argument for not giving the longest roads as much weight as their size would imply, and therefore that there could be some argument for using per-mile observations as a proxy for weighting inversely to size. On the other hand, dividing by miles is approximately equivalent to weighting by the reciprocal of miles squared ($1/M^2$), and that may be too much of a good thing. Hence we shall report also on some computations assuming intermediate weighting schemes: (1) weighting the per-mile figures by miles, which undoes some of the implicit weighting, resulting in an approximate weighting by the reciprocal of miles rather than of miles squared ($W = M/M^2 = 1/M$); and (2) using the logarithms of cost and output (not per mile) as variables, thereby reducing the influence of the extremely large or small observations.[14]

The results are arranged in Table 1 in three groups: I. Total Sample, $N=97$; II. Small Roads, $N=45$; and III. Larger Roads, $N=52$. The last group is the most relevant, in my opinion, but the others are also reported for comparison purposes and for the sake of completeness. Within each group, the results are arranged in order of decreasing weight given to smaller roads: (1) deflated (per-mile) units—implicit weighting of different observations approximately $1/M^2$; (2) deflated weighted by miles—approximate weighting by $1/M$; (3) logarithmic transformation (comparable very roughly to a weighting by the reciprocal of the square root of the variables); and (4) undeflated—and hence unweighted observations (or weighted proportionately to size). Since the argument for any deflation must be made on efficiency grounds, it is not unreasonable, other things being equal, to prefer that procedure which yields the highest precision (lowest standard error) in estimating the parameter of interest (percent variable).

Several things stand out in this table. First, the two size subgroups behave quite differently, have significantly different fits, different shaped lines, and different elasticities. Thus, combining both groups is inappropriate and leads to a mixture which depends crucially on what weight is implicitly assigned to the smaller roads. Similarly, it is inappropriate to estimate one logarithmic relationship for the whole sample, imposing a constant elasticity on two halves having very significantly different elasticities. Also, the estimated percent variable rises as less weight is given to the smaller roads, implying that if there are economies of scale, they are much less pronounced for the larger roads. Concentrating on the relevant larger size subgroup III., which accounts for about 95 percent of total railroad costs, the best estimates on the basis of relative fit and the precision with which the "percent variable" is estimated are the last two in the table, or 90 and 99.[15]

---

[14] A logarithmic transformation compresses the scale in which variables are measured, reducing a ten-fold difference between two numbers to a two-fold difference. It does, however, also affect the shape of the assumed relationship, implying a constant elasticity (percent variable) at all levels of output.

[15] The first figure is not strictly comparable since it assumes a curvilinear relationship between cost and output.

TABLE 1

RAILWAY COST FUNCTIONS,1957–1961 AVERAGES

SIZE GROUP AND ESTIMATION PROCEDURE		CONSTANT ("FIXED COST") a (DOLLARS PER MILE OR THOUSANDS OF DOLLARS)	SLOPE (MARGINAL COST) b (MILLS PER TON OR TON-MILE)	"AVERAGE" AVE COST $\bar{c}/\bar{x}$ $(\overline{C/M}) / (\overline{X/M})$	ELASTICITY "PERCENT VARIABLE"	R^2
I TOTAL, N=97	1. DEFLATED, PER–MILE UNITS	13,010 (6,218)	6.43 (0.87)	8.62	74.6 (11.3)	0.365
	2 DEFLATED WEIGHTED	3,805 (3,713)	6.06 (0.51)		92.6	0.826
	3. LOGARITHMS OF UNDEFLATED UNITS				82.3 (2.2)	0.944
	4. UNDEFLATED, ORIGINAL UNITS	2,811 (4,524)	6.39 (0.18)	6.59	97.0 (4.7)	0.944
II SMALL, N=45	1. DEFLATED	19,450 (9,927)	7.32 (1.36)	10.92	67.1 (11.9)	0.402
	2. DEFLATED WEIGHTED	15,450 (8,937)	6.24 (1.28)		73.8	0.394
	3 LOGARITHMIC				60.3 (6.7)	0.649
	4 UNDEFLATED	3,197 (1,962)	6.84 (1.19)	9.73	70.3 (15.4)	0.434
III LARGER, N= 52	1. DEFLATED	10,120 (5,181)	5.36 (0.74)	6.94	77.3 (11.2)	0.513
	2. DEFLATED WEIGHTED	2,702 (4,233)	6.10 (0.51)		93.9	0.956
	3. LOGARITHMIC				90.4 (2.7)	0.956
	4. UNDEFLATED	1,473 (9,718)	6.42 (0.25)	6.47	99.1 (5.7)	0.929

I TOTAL SAMPLE FOR WHICH OBSERVATIONS WERE AVAILABLE IN ALL FIVE YEARS

II. ROADS WITH LESS THAN 500 MILES OF TRACK OPERATED IN 1957

III ROADS WITH MORE THAT 500 MILES OF TRACK OPERATED IN 1957

1. EQUATION FORM $\frac{C}{M} = a + b\frac{X}{M}$, UNWEIGHTED.

2. $\frac{C}{M} = a + b\,\frac{X}{M}$, WEIGHTED BY M.

3. LOG C = LOG a+b LOG X, UNWEIGHTED.

4 C = a + b X, UNWEIGHTED

C–TOTAL COST PER RAILROAD, X–GROSS TON–MILES, M–MILES OF ROAD OPERATED. DATA TAKEN FROM COST SECTION WORKSHEETS

PERCENT VARIABLE OR THE ELASTICITY OF TOTAL COSTS WITH RESPECT TO A CHANGE IN OUTPUT COMPUTED FOR EQUATION.

1. $\text{p.v.} = b\,\frac{(\overline{X/M})}{(\overline{C/M})} = \frac{\text{MARGINAL COST}}{\text{AVERAGE COST}} = \frac{\text{AVERAGE COST PER MILE} - a}{\text{AVERAGE COST PER MILE}}$

2 $\text{p v} = \frac{(\overline{C/M}) - a}{(\overline{C/M})}$, THE WEIGHTED REGRESSION PROGRAM USED DID NOT CALCULATE THE CORRESPONDING WEIGHTED AVERAGES. INSTEAD WE USE THE COST SECTION FORMULA, SUBSTRACTING THE NEW ESTIMATE OF 'a' FROM THE UNWEIGHTED AVERAGE COST PER MILE AS IN 1. THIS IS NOT STRICTLY CORRECT, SINCE THE WEIGHTED LINE DOES NOT NECESSARILY GO THROUGH THIS MEAN, BUT IT IS CONSISTENT WITH THE COST SECTION PROCEDURES. ALTERNATIVELY, ONE CAN COMPUTE THE PERCENT VARIABLE BY USING THE AVERAGE COST AS ESTIMATED IN 3 FOR THE UNDEFLATED VARIABLES. AGAIN, THIS IS NOT STRICTLY COMPARABLE SINCE THESE AVERAGES ARE UNWEIGHTED, BUT THE DIFFERENCES ARE MINOR THIS ALTERNATIVE PROCEDURE YIELDS THE FOLLOWING PERCENT VARIABLE ESTIMATES I 92.0, II 70.3, AND III 94.3, RESPECTIVELY

3. p.v. = b, THE ESTIMATED COEFFICIENT IN THE LOGARITHMIC EQUATION, SINCE $b = \frac{d\,\text{LOG}\,C}{d\,\text{LOG}\,X} = \frac{dC}{dx} \cdot \frac{X}{C}$ AND IS CONSTANT FOR ALL POINTS ALONG THE EQUATION

4 $\text{p.v.} = b\,\text{TIMES}\,\frac{\text{AVERAGE GROSS TON MILES}}{\text{AVERAGE TOTAL COST}} = \frac{\text{MARGINAL COST}}{\text{AVERAGE COST}}$.

THE NUMBERS IN PARENTHESES ARE THE ESTIMATED STANDARD ERRORS OF THE RESPECTIVE COEFFICIENTS. THE STANDARD ERROR OF THE PERCENT VARIABLE IS APPROXIMATED, IN THE LINEAR CASE BY $\text{s.e.}(\text{p.v}) = \text{s.e}\,(b\bar{x}/\bar{y}) = \bar{x}\,\text{s e}\,(b/\bar{y})$, WHERE y IS THE PARTICULAR DEPENDENT VARIABLE AND x IS ASSUMED TO BE A FIXED NON–RANDOM VARIATE SINCE THE ESTIMATED b AND $\bar{y}$ ARE INDEPENDENT, WE HAVE

$$\text{s e.}(\text{p v.}) = \bar{x}\,\text{s.e}\,(b/\bar{y}) \simeq \bar{x}\,\frac{b}{\bar{y}}\sqrt{\frac{\sigma_b^2}{b^2} + \frac{\sigma_{\bar{y}}^2}{\bar{y}^2}}$$

WHICH CAN BE SIMPLIFIED FOR COMPUTATIONAL PURPOSES TO

$$\text{s e p v} \simeq \text{p.v.}\sqrt{\frac{1}{t_b^2} + \frac{(1-R^2)}{N}\left(\frac{\sigma_y}{\bar{y}}\right)^2}$$

WHERE $t_b = b/\sigma_b$

R^2 IS SQUARE OF THE MULTIPLE CORRELATION COEFFICIENT

Note, again, that there is much less difference in the estimated marginal cost or slope coefficients, and that most of the difference in the various estimates of "percent variable" are due to differences in the associated average cost. But the weighting appropriate for computing the "average" relevant for policy purposes is not the same as the weighting which may be optimal in minimizing the errors of estimation.

Eliminating the very small roads from consideration and assuming that the cost function is linear, one cannot reject the hypothesis that there are no economies of scale at all (that p.v. is 100). In the linear case, the significance of economies of scale or decreasing costs depends on whether the estimated constant a is significantly different from zero. The conventional level of testing such hypotheses is "two standard errors" or the 95-percent confidence level, implying roughly that one would not reject a hypothesis if the observation we are examining could arise in 5 percent or more of the samples if our hypothesis were true. That is, given the hypothesis, the observation is not a "rare" enough event to contradict it. Using this criterion one cannot dismiss the no increasing returns hypothesis even for the deflated estimates, since the estimated "fixed costs" of $10,120 are less than twice their estimated standard error of $5,181. For the weighted and undeflated estimates, the estimated a is practically undistinguishable from zero since its estimated standard error (s.e.) is as large or even larger. In the logarithmic equation we cannot reject the hypothesis that the increasing returns are not larger than 4 percent or that the true percent variable is 96.

Some of these findings are illustrated in Figure 5, which plots on double logarithmic paper the five-year averages of total costs and gross ton-miles for each railroad. On such paper the percent variable can be read directly as a slope of the line through a set of points. For the percent variable to be a constant this line should have the same slope, i.e., be a straight line on this paper. I have drawn three lines through these points. The first is a line having a slope of unity drawn through the mean of all of the points. If all of the points lay approximately along such a line, the percent variable would be approximately 100 for all points. This line is drawn to facilitate a visual comparison of the amount of deviation of the actual relationship between these points from such a line. In addition, I have drawn "best fit" lines for the larger and smaller groups of roads, separately. These lines, as noted in the table, have slopes of 0.904 and 0.603.

4. Related studies

■ The finding that there is very little or almost no evidence of economies of scale in the railroad industry (if one excludes the smallest roads from consideration) is consistent with the findings of the few published econometric studies of this subject.

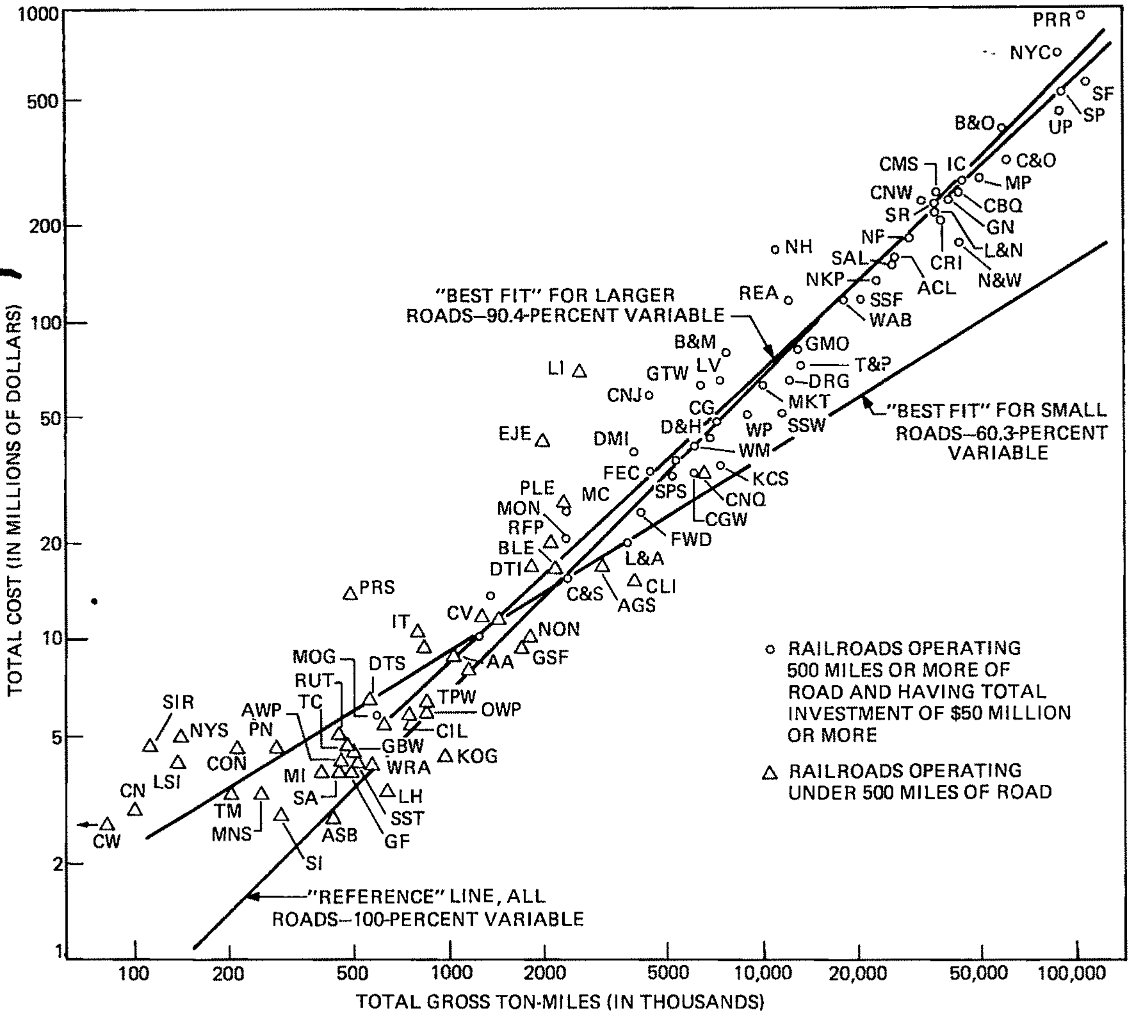
As far as I am aware, there are only four modern econometric studies of returns to scale in the railroad industry: two studies by Professor George H. Borts, one by Professor Lawrence R. Klein, and a book by J. R. Meyer et al.[16] Since the last work does not focus on the question of economies of scale directly and does not provide the additional numbers required to derive the implica-

---

[16] See Borts [2,4]; Klein [6], pp. 226–36; and Meyer et al. [8]. See also Borts [1].

FIGURE 5

TOTAL COST RELATED (LOGARITHMICALLY) TO TOTAL GROSS TON-MILES
ALL U S CLASS I RAILROADS–AVERAGE 1957-1961



SOURCE·ICC COST SECTION–WORK TABLE 2.

tions of their findings for this topic, it will not be discussed further here. In his 1952 *Econometrica* study,[17] Borts investigates the relationship of inputs to outputs in the railroad industry. He divides the railroad freight operations into two broad categories: the switching process and the line-haul process. The variable input variables are the same for both categories: labor in man-hours, fuel consumed in net tons, equipment maintenance in dollars, and track and structure maintenance in dollars. In addition, he considers "unadjustable" size-related inputs—total tractive capacity of locomotives and miles of main line track for the line-haul process, miles of switching yard track, tractive capacity of yard locomotives, and the average number of freight cars standing on the line for the switching process. The output dimensions considered are loaded freight car miles, carloads of freight, and empty freight car miles for the line-haul process, and yard switching locomotive miles, yard switching locomotive hours, and carloads of freight for the switching process. The statistical procedure used is to estimate a separate equation for each input and each process connecting the input (cost) level to the output level and size (unadjustable inputs) variables, using data for 76 class I railroads in 1948. These equations are estimated in both deflated (by reproduction costs and yard miles) and undeflated form, with the deflated form resulting in somewhat poorer and less stable estimates. Using his estimates, he computed separate percent variable estimates for each input category and process and a weighted average percent variable (elasticity of variable cost). This overall estimate is 94 with a standard error (s.e.) of 10 for the line-haul process and 103 (s.e. = 17) for the switching process.[18]

---

[17] See Borts [4].

In his 1960 *Econometrica* article,[19] Borts relates allocated freight operating expenditures to total loaded and empty freight car miles and total freight carload loads. This study is notable for the very ingenious, complex, and detailed attempt to allow for size and regional differences in the estimates of the overall cost function. The function is estimated, essentially separately, for three regions and three size classes and using two alternative size deflators—reproduction costs and track mileage. The first deflator gives somewhat better results in terms of fit and t-ratios for the coefficients. Cost elasticities are computed separately for each region and size class, but the best estimates are those computed "between size classes," since they are least subject to downward bias due to errors of measurement and short-run influences (the "regression fallacy"). These vary from 97 to 140 for the Eastern region, from 83 to 100 for the Southern region, and from 93 to 102 in the West. Also, the "within size class" estimates are significantly higher for the largest size class. In fact, Borts concludes that "*increasing* costs apparently prevail in the East." His findings are consistent, however, with some, but not very large, decreasing costs in the South and West.

In his econometrics textbook,[20] L. R. Klein reports on a larger but never published study of "Production of Railway Services." The model used is quite elaborate, relating input categories of man-hours, tons of fuel, and train hours to the prices of labor, fuel, and capital, and the output levels of freight (net ton-miles) and passenger (net passenger miles) services. The prices are used to estimate the minimum total variable cost levels for each railroad, and the resultant construct is related to the two output variables. His numerical results can be transformed into an equation which allows one to estimate the total percent variable, which, given his estimates, is 90 percent.[21]

---

[18] These are from the better fitting undeflated equations. In deflated form the respective p.v.'s are 55 (12) and 99 (10).

[19] See Borts [2].

[20] See Klein [6].

[21] His final equation (p. 234) is

$$N^{1\ 0} C^{135} D^{312} = A X_1^{1\ 122} X_2^{181},$$

where N, C, and D are labor, fuel, and capital (train hours) inputs, and X_1 and X_2 are net freight ton-miles and net passenger miles, respectively. This equation can be rewritten as

$$[N^{69} C^{09} D^{22}] = k[X_1^{86} X_2^{14}]^{90},$$

where the terms in the brackets are a total input index and a total output index, respectively, with within index weights (coefficients) adding up to unity.

Similarly, the undeflated recomputations of the Cost Section data for the separate years 1957–1961 performed by Professor Glasser imply percent variable estimates between 95 and 100.[22] While it is tempting to conclude that since he gets essentially the same results for five separate years as we do for the average of these five years, these findings reinforce each other and the true variability around them is smaller than is indicated by the computed standard errors; but this is false. It would be true only if each of these years represented an independent sample; but this is far from true. The relative cost and output relationships change very little from year to year and hence provide almost no new information. In fact, the correlation between gross ton-miles per mile in 1960 and gross ton-miles per mile in 1961 is 0.999. Similarly, the correlation between the residuals from the per-mile regressions in these years, which are supposed to be estimates of the random components, is also 0.999. Thus virtually the same information is contained in each of these years, and hence while this leads to similar results for the separate years, these results cannot be interpreted as supporting each other; they are the same results. Moreover, such a high correlation among what purports to be random components implies the existence of systematic forces affecting the cost-output relationship which have been left out of the cost function formulation developed by the Cost Section.

5. Conclusions

■ We have seen that even if one accepts the definitions of cost and output underlying the Cost Section study, the Section's results are extremely sensitive to the choice of particular observations and to the statistical procedure used. In particular, the Cost Section procedure gives inordinate weight to the experience of the very small railroads in estimating the "average" percent variable. Moreover, we have shown that miles of track is a poor size variable and hence the per-mile based results are inefficient and probably also biased. When the smallest roads are eliminated from the sample, and when the cost function is not divided by the irrelevant miles variable, the estimated percent variable is substantially higher than the 80-percent Cost Section figure. There is not much difference between the various estimates of marginal cost (the slope of the function), but the estimate of average cost, and hence also of the elasticity (percent variable), depends very much on the grouping and weighting of the observations to be averaged. At the "average" levels of the country as a whole, with each ton getting equal weight in this concept, average costs are much lower than those estimated by averaging the separate per-ton-mile figures for each road and giving equal weight to roads rather than tons. At the relevant average the percent variable appears to be above 95.

These conclusions, however, are based on very questionable definitions of cost and output, and on a very gross aggregation of types of traffic, dimensions of output, regions of the country, and sizes of railroads. Nevertheless, more sophisticated econometric studies which allow for some of these complications also reach the conclusion that the "true" percent variable is not much less than

---

[22] See Docket No. 34013 [11], Vol. II of the submission by the motor carriers, Exhibit D, April 1965.

100. If there are some decreasing costs, they are not prevalent in the East or among the bigger railroads, and they are nowhere near as large as is implied by the Cost Section's 80-percent figure.

In a way this finding is not as surprising as it sounds. There may well be decreasing average costs for some types of traffic, at some times, in some areas. But all of the studies examined ask the question, What will happen to average costs if total traffic is expanded on the average in the same proportions and having exactly the same distribution over the various commodities, types, routes, and seasons as the previously handled traffic? There may be very little return to scale from a proportionate increase in all kinds of traffic. Whatever decreasing costs there may be are likely to arise only if one can contemplate disproportionate changes in traffic, changes in some kinds of traffic but not in others. But that cannot be discovered from such studies as we have examined above. It requires a different and much more *ad hoc* research program.

The Commission would also benefit from the pursuit of more detailed and sophisticated studies. The studies underlying the Cost Section recommendations are at least ten years behind the state of the art in statistical investigations of economic data. The range of tools of analysis available to economists has grown tremendously over the last two decades, both in the strength of these tools and their sophistication. The question of the existence of economies of scale is a difficult but an interesting one; substantial work has been devoted to it, and a number of important research results have been published.[23] The work of the Cost Section shows no awareness of the existence of such literature. Also, it is unlikely that any professional journal would accept for publication the results reported by the Section, even though most journals would be interested in publishing research results on this topic.[24] The Commission is in possession of a vast store of data supplied to it annually by the railroads. It can also ask for additional data if it finds that necessary. Thus, there is no real barrier to the pursuit of much richer and more relevant studies of this subject. The Commission, the railroads, and the taxpayer deserve something better.

Appendix

■ On July 13, 1970 the ICC in Docket No. 34013 made the following decisions:

(1) Approval and adoption of specific cost formulas, with a view toward giving prima facie validity to formula-based costs [were] found not shown to be necessary or desirable.

(2) The terms "out-of-pocket costs" and "fully distributed costs," as used in Commission proceedings, [were] required to be changed to "variable costs" and "fully allocated costs," respectively, and the noncost elements of profit, income taxes, and for railroads, the passenger and less-than-carload deficits, [were to be] excluded therefrom.

---

[23] For example, see Johnston [5].

[24] For example, the argument presented in the Day statement [12] against multiple regression techniques is both wrong and specious. A doubling of variables does not require a doubling of the number of observations. What matters are the "degrees of freedom." Besides, with the mounds of data it is sitting on, the Section should not have to worry about this problem.

(3) The determination of a variability factor for particular services [was] found to require the selection of an appropriate time period sufficiently long to reflect adequately those changes in operations resulting in expenses which can reasonably be expected to vary with the performance of the particular service or services rendered.

(4) The allocation of constant costs to particular services, for ratemaking purposes, result[ed] in the assignment of an equitable portion of such expenses to the particular services, and no single method [was] found universally applicable to all transportation services.

(5) No single method of apportioning joint or common costs [was] found universally acceptable, and any method of apportionment utilized for ratemaking purposes [was] required to be designed to reasonably reflect the specific circumstances attending the transportation performed.

(6) In appropriate circumstances, (a) "fully allocated costs" [were] found representative of the full expense level assignable to particular services; (b) relevant "variable costs" [were] found indicative of the minimum level of expenses which must normally be recovered by a carrier in providing particular services; and (c) "incremental" or "marginal" costs [were] authorized for utilization as indicative of a minimum expense level for ratemaking purposes in appropriate short-run situations.

(7) A task force of Commission employees [was] authorized for the purposes of conferring with shipper, carrier, and government agency representatives to consider the feasibility and practicability of additional research projects into specific areas of transportation cost ascertainment.

References

1. BORTS, G. H. "Cost and Production Relations in the Railway Industry." Unpublished Ph.D. dissertation, University of Chicago, 1953.
2. ———. "The Estimation of Rail Cost Functions." *Econometrica*, Vol. 28, No. 1 (January 1960), pp. 108–31.
3. ———. "Increasing Returns in the Railway Industry." *Journal of Political Economy*, Vol. 62, No. 4 (August 1954), pp. 316–33.
4. ———. "Production Relations in the Railway Industry." *Econometrica*, Vol. 20, No. 1 (January 1952), pp. 71–79.
5. JOHNSTON, J. *Statistical Cost Analysis* New York: McGraw-Hill Book Co., 1960.
6. KLEIN, L. R. *A Textbook of Econometrics.*.Evanston, Ill.: Row, Peterson & Co., 1953.
7. KUH, E. and MEYER, J. R. "Correlation and Regression Estimates when Data are Ratios." *Econometrica*, Vol. 23, No. 4 (October 1955), pp 400–16.
8. MEYER, J. R., PECK, M. J., STENASON, J., and ZWICK, C. *The Economics of Competition in the Transportation Industries.* Cambridge, Mass.: Harvard Univ. Press, 1959.
9. U. S. BUREAU OF THE CENSUS. *Statistical Abstract of the U. S.: 1960.* 81st ed. Washington, D. C.: U. S. Government Printing Office, 1960.
10. U. S. INTERSTATE COMMERCE COMMISSION, BUREAU OF ACCOUNTS. *Explanation of Rail Cost Finding Procedures and Principles Relating to the Use of Costs.* St. 7–63. Washington, D. C., November 1963.
11. ———. "Rules to Govern the Assembling and Presenting of Cost Evidence." Docket No. 34013, 321 I.C.C. 238, Order of April 16, 1962.
12. ———. Statement of H. J. Day, in Docket No. 34013 [11], September 1964.

Antitrust laws and public utility regulation

Phillip E. Areeda

Professor of Law
Harvard University

Many of those concerned with public utility regulation assume that federal antitrust law has little bearing on regulated industries. That assumption is erroneous. This paper indicates some of the ways in which antitrust law affects regulated firms. It begins with a few words about antitrust premises and characteristics of regulation and then turns to three topics: antitrust principles as applied by federal regulatory agencies, antitrust law as applied by the courts to federally regulated firms, and the interface of federal antitrust law with regulation by the several states.

1. Introduction: an antitrust primer

■ To understand the intersection of federal antitrust law with direct regulation by federal or state agencies, it is important to have in mind several fundamental characteristics of antitrust law and its implementing institutions. These are intended to protect competition from the business behavior or market structures that endanger competitive processes or threaten to deny society the price, output, and innovative benefits toward which "workable competition" tends. To be sure, antitrust law is sometimes applied to protect existing firms from competition, as some notorious applications of the Robinson-Patman Act's prohibition of price discrimination seem to suggest. And the judges sometimes seem to say that the Sherman and Clayton Acts are intended to preserve a world of numerous small producers regardless of efficiency. Yet the few recent merger cases speaking in these terms actually exhibited an implausible or minor claim of efficiency or no such claim by the merging parties at all.[1] The important point is that anticompetitive applications and language are the exceptions, ripples on an antitrust mainstream protecting and promoting that competition which is feasible and practicable.

The key difficulty for antitrust agencies is the claim, in a particular case or class of cases, that competition is not possible, not practicable, or not fully responsive to the social interest. Such claims are properly viewed with skepticism, for there has never been an antitrust sinner who did not claim that he was serving the public good. But the claim is sometimes true, and an obvious saving in social resources can effectively immunize what would otherwise be

---

Phillip E. Areeda received the A.B. and L.L.B. degrees from Harvard University (1951 and 1954, respectively). He is presently a faculty member of the John F. Kennedy School of Government at Harvard, as well as of the Harvard Law School. From 1956–1961 he was Assistant to the Special Counsel to President Eisenhower, and in 1969, the Executive Director of the Cabinet Task Force on Oil Import Control. Professor Areeda is the author of a book on antitrust analysis.

[1] FTC v. Procter & Gamble Co., 386 U. S. 568 (1967); Brown Shoe Co. v. United States, 370 U. S. 294 (1962); United States v. Bethlehem Steel Corp., 168 F. Supp. 576 (S.D.N.Y. 1958); United States v. Third National Bank in Nashville, 390 U. S. 171 (1968).

an unlawful restraint of trade. Even restraints said to be "per se" unlawful can be validated, at least for a time, when necessary for the achievement of a legitimate object, such as a productive joint venture, and when competition is not otherwise possible.[2] The difficulty lies in deciding what constitutes a legitimate object and whether the challenged behavior is truly necessary to achieve it. In short, the Antitrust Division, the Federal Trade Commission, and the courts face the unenviable task of distinguishing good from bad when both are indistinct. The possible social harm might be qualitatively significant or not, likely to occur or not, and if it does occur, quantitatively significant or not. And although the courts will not lightly forbid the parties from achieving legitimate objectives, there are the usual elusive issues of fact and judgment. A claimed objective may be obviously desirable, questionably desirable, or not desirable at all. It may flow obviously from the challenged conduct or only doubtfully so. It may or may not be significant in magnitude. And it may or may not be achievable in less restrictive ways presenting fewer social dangers.

The essential fact of life—for regulation as well as for antitrust—is that data about the past, predictions about the future, and judgments about anything are not always extant, costless, reasonably obtainable, or very reliable. Furthermore, decision-makers everywhere are of mixed quality and sometimes of no quality at all.

2. Government intervention

■ To speak, as we have, of protecting market processes is not a call for laissez-faire, for it is silly to rely on the market for that which it cannot do. We may make collective value judgments through government specifications or collective consumption. Where the market does not reflect external costs or benefits, the law may impose the costs, grant bounties, or supplant the market altogether. It is important to remember, however, that comprehensive, full-blown, public-utility type regulation may be unnecessary to achieve a stated public objective. There is no need to control the number of entrants or their prices or output in order to achieve the purity of drugs; the safety of airplanes, taxicabs or trucks; security for bank depositors; minimum competence for lawyers, doctors, or plumbers; or honest or efficient stockbrokers. Unfortunately, however, regulation may be adopted in order to suppress competition that is possible and practicable. And whatever its original purpose, regulation sometimes proceeds with its own dynamic.[3] Natural monopoly in local electric or telephone service does not dictate monopoly in the manufacture of electric or telephone equipment or in the supply of other energy sources or communication services. One aspect of this point can be made in another way: The regulator's power and/or its exercise is never complete. Varying degrees of discretion are always left to the management of the regulated enterprise, and that management may act in ways that impair such competition as remains possible, notwithstanding the existence of regulation.

---

[2] United States v. Pan American World Airways, Inc., 193 F. Supp. 18 (S.D.N.Y. 1961), reversed on other grounds, 371 U. S. 296 (1963).

[3] Readers of this journal will recall comments on this tendency in articles by Professors McKie, Stigler, and Posner in the Spring 1970 and Spring 1971 issues. See J. W. McKie, "Regulation and the Free Market: the Problem of Boundaries," *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 6–26; G. J. Stigler, "The Theory of Economic Regulation," *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 3–21; and R. A. Posner, "Taxation by Regulation," *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 22–50.

3. Antitrust principles as applied by regulatory agencies

■ **Why speak of "antitrust" principles?** When a regulatory agency examines a cartel agreement, a merger, or the buying practices of a regulated firm, it necessarily considers the kind of issues considered by a court faced with similar practices by unregulated firms. The regulatory situation might, of course, call for a different resolution, but the principles applied are similar. To be sure, antitrust law can claim no monopoly on the use of competitive concepts. Yet, it is appropriate to speak of "antitrust" principles for reasons of history and of judicial review.

It is in the antitrust context that competitive concepts have been most intensively explored by governmental institutions.[4] Second, the work of federal agencies is subject to judicial review, and the courts will not enforce the agency decision unless the judges are satisfied that the agency has considered the relevant issues presented by the case at hand. The judges' own understanding of competitive issues is mainly derived from antitrust cases. To show the court that it has done its job, therefore, the agency may feel obliged to discuss competitive issues in the antitrust terms familiar to the courts.

☐ **Antitrust claims by parties or intervenors.** There are few limitations on the persons who may appear before an agency to make a vague and conclusionary claim that a regulated firm's acts or proposed acts will have an unreasonable competitive effect. Governing statutes often grant any "person aggrieved" standing to raise questions before the agency and then to challenge the agency action in court. A competitive firm or industry or even a consumer as a "private attorney general" may be granted standing.[5] And not only is the range of persons broad, the issues they are permitted to raise sometimes seem limitless.[6]

---

[4] Many economists think that no governmental institution has analyzed competitive principles very far, but that is another essay.

[5] Hardin v. Kentucky Utilities Co., 390 U. S. 1 (1968); FCC v. Sanders Bros. Radio Station, 309 U. S. 470 (1940); City of Pittsburgh v. FPC, 237 F.2d 741 (D. C. Cir. 1956); Associated Industries v. Ickes, 134 F.2d 694 (2d Cir. 1943), vacated for mootness, 320 U. S. 707 (1943).

[6] As a possible extreme case, consider City of Pittsburgh v. FPC, 237 F.2d 741 (D. C. Cir. 1956). A natural gas pipeline sought, as it must, FPC permission to abandon an unnecessary pipeline which it planned to convert to oil carriage. The FPC has no jurisdiction over oil pipelines as such. Once the line was converted to oil, it would compete, allegedly unfairly, with barges which were then carrying petroleum products to points that could be served by the converted pipelines. The barge operators insisted that the FPC abandonment proceedings weigh the effect of conversion upon them. The Commission declined, saying that it had neither power over nor interest in barges or oil pipelines which were not within its jurisdiction. The Court of Appeals reversed, holding that the Commission did have power to consider the question and was apparently obligated to do so because the statutory standard governing FPC decisions is whether the "public interest" would be served. The state of competition in petroleum carriage is relevant to the

□ **Competition as in "the public interest."** The antitrust laws are understood to express a national policy favoring competition, and that policy must be given appropriate weight by the administrators even though they have the express statutory power to grant an exemption from the antitrust laws. This proposition is illustrated by the *McLean* case,[7] where Interstate Commerce Commission approval—and consequent antitrust exemption—of a merger of truckers was challenged on the ground that the Commission had failed to apply the antitrust law standard. The Supreme Court rejected that challenge and upheld the Commission. But although the Commission decision is not precisely controlled by the standard of antitrust law, the Court made it clear that the Commission is obliged to consider the effect of the merger on competitors and on the general competitive situation in the light of the objectives of the regulatory statute.

The general notion that the regulator ought to value competition to the extent consistent with regulatory goals will provide clear guidance where the regulatory premise is both clear and inapplicable to the challenged conduct. For example, the belief that shipping service by regularly scheduled liners justifies the toleration of shipping cartels is no reason for allowing the cartel to boycott travel agents who also book passengers on non-cartel ships.[8] The limited broadcast spectrum must be allocated somehow. But the necessity of preventing physical interference among freely-entering broadcasters does not warrant limiting cable TV operations in order to protect those who do receive a license.[9] And even if telephone service were considered a natural monopoly, equipment manufacture is not. There is, therefore, little warrant for agency approval of a telephone company tariff preventing customer installation of auxiliary devices that do not impair the general system; there are less restrictive ways of requiring compliance with reasonable technical standards designed to insure "system integrity."

The proper place for competitive standards is more difficult to locate when the legislative mandate is understood to include promotional or redistributive objectives. Are airlines regulated in order to give the public the benefits of competition in the absence of perfectly competitive market structure? Or does regulation seek the

---

public interest Therefore, the FPC must consider and hear evidence on the barge operator's claim. The court apparently gave no weight to the fact that the FPC faced an enormous backlog of uncompleted work lying within its core responsibilities. It seems obviously unwise, however, to allow intervenors with quite remote interests to control the agency's allocation of its limited resources. One does not wish to see intervenors with a selfish or even anticompetitive interest control the agency's timetable. Nor is one altogether happy about permitting passionate members of the general public to do so. There is, nevertheless, a genuine dilemma because the regulatory agencies themselves have sometimes demonstrated a lax insensitivity to the general interest. But unless undue resources are to be given to pursuing tangential claims—exhausting agency resources and delaying the satisfaction of public needs—the courts will have to distinguish the more important from the less important in requiring agency attention to the claims of intervenors or the general public.

[7] McLean Trucking Co. v. United States, 321 U. S. 67 (1944).

[8] Federal Maritime Commission v. Swedish-American Line, 390 U. S. 238 (1968).

[9] Although rationing is the *raison d'etre* for FCC licensing, the Commission may believe that it has the duty to assure free over-the-air TV and thus to protect at least some broadcasters from possible ruination at the hands of cable systems.

development of more elaborate equipment through cartel pricing that diverts competitive impulses into service and equipment variations? Or is the purpose perhaps to hold up rates between major markets in order to provide sufficient carrier revenues to subsidize non-self-supporting service to and among lesser markets—possibly to further national cohesion or industrial and population dispersal?

In yet another class of cases, the regulatory and antitrust objectives are similar: as much competition as is practicable, consistent with the economic and technical imperatives of efficiency. This objective guides both antitrust and administrative doctrines about mergers, but the implementing approaches have developed differently. With respect to horizontal mergers, antitrust courts have come to adopt a rather simple prohibition of mergers involving an undue market share[10]—often overly simplistic and sometimes based on gerrymandered definitions of the market.[11] In contrast, the regulatory agencies purport to consider every factor bearing on competition and efficiency. Yet it must be said that railroad merger decisions over the past 15 years evidence cumbersome agency and judicial procedures burdening the parties and the government without illuminating the issues. In so many of the decisions, the anticompetitive tendencies and the alleged offsetting cost savings were usually addressed at a level of vague generality that was quite unpersuasive. And widely remarked have been the decisions of some Comptrollers of the Currency who approved bank mergers with no apparent justification beyond the parties' desires. In many instances it may be that the critical difference between courts and administrators is one of methodology, as the different approaches toward mergers illustrate.

Often, however, the difference will be one of attitude. Usually, the antitrust authorities will be more skeptical about firm or industry claims of social efficiency. An interesting and apparent reversal of outlook seems to be illustrated by the Federal Maritime Commission. Its predecessors, in days past, approved shipping cartel practices injuring non-cartel ships without inquiring into the necessity of such coercive tactics to achieve the purposes of the statute. More recently, the Commission condemned a cartel of Atlantic passenger carriers who boycotted travel agents booking any passengers on non-cartel ships. The Commission disapproved of this practice, in the words of the Supreme Court, on a "principle that conference restraints which interfere with the policies of antitrust laws will be approved only if the conferences can 'bring forth such facts as would demonstrate that the . . . rule was required by a serious transportation need, necessary to secure important public benefits or in furtherance of a valid regulatory purpose of the Shipping Act.' " The Supreme Court upheld the agency's refusal to grant antitrust immunity to the cartel. The Court held "that the antitrust test formulated by the Commission is an appropriate refinement of the statutory 'public interest' standard."[12]

---

[10] E.g., United States v. Von's Grocery Co., 384 U. S. 270 (1966).

[11] E.g., United States v. Aluminum Co. of America (Rome Cable), 377 U. S. 271 (1964); United States v. Continental Can Co., 378 U. S. 441 (1964).

[12] Federal Maritime Commission v. Swedish American Line, 390 U. S. 238, 243, 246 (1968).

4. Antitrust law applied by the courts to federally regulated firms

■ To understand the antitrust court's relation to the federally regulated industry, we turn to that lawyer's mystery known as "primary jurisdiction."

☐ **The shipping conference case.**[13] Cartel agreements specifically approved by the Federal Maritime Commission were expressly exempted from federal antitrust law. Failure to file such an agreement violated the Shipping Act, and such violations could be punished by penalties of $1,000 per day and were also subject to "cease and desist" orders from the Maritime Commission. The *Far East* case involved an unfiled cartel agreement that included a so-called dual rate provision for higher charges to shippers not agreeing to use cartel vessels exclusively. The Antitrust Division attacked the cartel in court and sought an injunction against continuation of the cartel agreement that was unfiled and therefore necessarily not exempted by any Commission action. Nevertheless, the Supreme Court dismissed the complaint.

In dealing with such an industry, perhaps one should hold that the existence of regulation supersedes antitrust law, such that the Commission would become the exclusive judge and enforcer of the competitive rules of the game. To hold that regulation would oust judicially enforced antitrust law has the virtue of unencumbered simplicity. But total displacement would also have several costs. Improper conduct could no longer be challenged directly by injured private persons. To be sure, they could complain to the Commission, but they would have no independent remedy. Second, the flexible arsenal of antitrust remedies would be lost. In place of injunction, private damages, and criminal sanctions, there would be only the administrative cease and desist order and the $1,000 a day civil fine. Third, the general and comprehensive proscription of unreasonable agreements and/or anticompetitive exercises of power would be lost. In their place would be a rather more specific catalog of permitted and prohibited acts. Fourth, in place of impartial, disinterested, skeptical courts, private plaintiffs, and the Justice Department, decisions would be made by administrators who sometimes grow too close to their regulated clients. Finally, total displacement would seem inconsistent with the statute. Its specific antitrust exemption for approved agreements would imply a congressional intention that unapproved conduct remain subject to the antitrust law.

If one rejects the view that regulation supersedes antitrust, should one take the rather obvious view that an unfiled, unapproved, and unexempted cartel agreement is to be tested by ordinary antitrust standards? Of course, the basic antitrust standard of reasonableness takes into account the particular circumstances of the industry. But the agency charged with regulating the particular industry is presumably more knowledgeable and more expert with respect to that industry and thus better able to judge the reasonableness of competitive or anticompetitive behavior there.[14] The question, then, is this: Who is best able to make an intelligent judgment about market behavior, the agency or the court?

---

[13] Far East Conference v. United States, 342 U. S. 570, 574–575 (1952).

[14] To be sure, the defendants could have availed themselves of administrative expertness through the simple expedient of obeying the Shipping Act and filing their agreement. Their failure to file deprives them of our sympathy. But that is not really the issue.

It was this issue of relative competence that the Court emphasized in the *Far East* case: "[C]ases raising issues of fact not within the conventional experience of judges, or cases requiring the exercise of administrative discretion" should be decided first by the relevant administrative agency, "even though the facts after they have been appraised by specialized competence serve as a premise for legal consequences to be judicially defined." And further, "Uniformity and consistency in the regulation of business entrusted to a particular agency are secured, and the limited functions of review by the judiciary are more rationally exercised, by preliminary resort for ascertaining and interpreting the circumstances underlying legal issues to agencies that are better equipped than courts by specialization, by insight gained through experience, and by more flexible procedure."[15] It should be observed that this formulation does not necessarily oust the antitrust laws. It merely says that the agency must be consulted first. And once the agency is consulted, the courts are available for antitrust remedies that the agency could not grant, such as criminal penalties or private treble damages. Such damages were allowed in a subsequent case after the Commission had in fact held that the defendant's price-fixing agreement was unfiled, unapproved, and unlawful.[16]

□ **The *Pan American* case.**[17] At the request of the Civil Aeronautics Board, the Justice Department sued Pan American, Grace, and their jointly-owned subsidiary, Panagra, alleging antitrust violations in (1) the 1929 formation by Pan Am and Grace of the joint subsidiary, Panagra, (2) Pan Am's use of its ownership interest in Panagra to prevent the latter from applying to the CAB for new routes competing with Pan Am, and (3) a Pan Am-Panagra agreement dividing their South American operations into noncompetitive segments. The alleged behavior began long before enactment of what is now known as the Federal Aviation Act. That Act requires filing of agreements among carriers and exempts those approved by the Board from the antitrust laws. The Board is also authorized to prohibit "unfair or deceptive practices or unfair methods of competition."

The Supreme Court dismissed the antitrust complaint, noting that the Board can immunize conduct from the antitrust laws and can halt all unfair practices and unfair methods of competition, including those which started prior to the Act. The Court went on to declare that "if the Courts were to intrude independently with their construction of the antitrust laws, two regimes might collide." Accordingly, the Court held that all questions of injunctive relief had to be left to the Board, but also declared itself hesitant to hold that a regulatory scheme "was designed completely to displace the antitrust laws, absent an unequivocally declared Congressional purpose to do so." For example, "the whole criminal law enforcement problem remains unaffected by the Act."

---

[15] It is ironic that, in a later episode, the Commission's approval of dual-rate cartel provisions was reviewed by the unspecialized courts and held inconsistent with the Shipping Act: Federal Maritime Board v. Isbrandtsen Co., 356 U. S. 481 (1957). Although Congress subsequently amended the statute to empower the Commission to approve dual-rate agreements, the statute imposed significant limitations on such cartels beyond anything the Commission had earlier done.

[16] Carnation Co. v. Pacific Westbound Conference, 383 U. S. 231 (1966).

[17] Pan American World Airways v. United States, 371 U. S. 296, 305, 310, 313 n.19 (1963).

Although the Court expressly left criminal antitrust remedies intact, and presumably private treble damage actions as well, it did not illuminate the respective roles of Board, judge, and jury in the course of such an action. The Supreme Court did note the availability, at least in civil cases, of a procedure whereby "litigation is held by a court until the basic facts and findings are first determined by the administrative agency." But is the agency to collect evidence or find facts on the context in which the parties operated—that is, what they did, alternatives open to them, the possibilities and costs of acting differently, and the social consequences of different actions? It may very well be that the agency could do these things better than a court. But the Court did not explain whether the judges would remain the ultimate arbiter of the appropriate degree of competition for the regulated industry. For the future, the same standard must govern court decisions as that which governs agency decisions. And even as to the past, the court cannot adopt an independent standard if the regulatory scheme is to be administered as a coherent whole.

☐ **The *El Paso* case.**[18] The Natural Gas Act required FPC approval before any regulated pipeline may "acquire or operate any . . . [gas] facilities." Some believed that Commission approval was required for asset acquisition but not for stock acquisition of another pipeline. One pipeline did acquire the stock of another. When the government sued to enjoin the merger under Clayton Act Sec. 7 the parties rearranged their transaction in the form of an asset acquisition and sought FPC approval. The District Court decided to delay its antitrust law decision to see whether the FPC approved. Commission disapproval would moot the antitrust issue; and even if the Commission approved, its investigation and reasoning might be helpful to the Court. The FPC did approve, and upon judicial review of that approval, the Supreme Court decided that such approval would not immunize the merger from the antitrust law. The Court also held, surprisingly and incomprehensibly, that the District Court had to decide the antitrust question before the Commission decided the regulatory question. Justice Douglas, writing for the Court, seemed to fear that the Commission's decision would influence the trial judge in some improper way.

The result is very curious. Obviously, the existence of pipeline regulation qualifies the applicability of the ordinary competitive premises embodied in the antitrust law. If a court is ultimately to decide the merger question, it must, of course, take account of the special circumstances of this industry. It is hard to see why the court should be compelled to make its decision without help from the regulatory agency. The antitrust court is likely to be adrift, cut off by the regulatory statute from its usual premises about the role of competition, and deprived by Justice Douglas of any FPC findings to inform its antitrust judgment. To be sure, some would doubt whether the FPC decision was, in fact, informative, but that is a different question.

---

[18] California v. FPC, 369 U. S. 482 (1962).

☐ **Bank mergers.** Banks are regulated in a variety of ways by a variety of federal and state agencies. There is no mechanism for administrative approval of agreements among banks, and so it is understandable that such agreements remain fully subject to normal antitrust law proscription. But administrative approval, often by several federal agencies, is a prerequisite to bank mergers. Agency approval, however, has little effect on the application of antitrust proscriptions to bank mergers.

In the *Philadelphia Bank* case,[19] the Supreme Court applied Clayton Act Sec. 7, the anti-merger provision, to a bank merger. Although this holding was incorrect, there was never any doubt in the legislative history of the 1960 Bank Merger Act that bank mergers were fully subject to the Sherman Act proscriptions against monopoly and unreasonable restraints of trade.[20] In any event, Congress reacted to the Philadelphia Bank holding by enacting the 1966 Bank Merger Act which (1) left Sherman Act Sec. 2 monopoly suits untouched, (2) limited other anti-merger suits against bank mergers to the 30-day period following administrative approval, and (3) modified the legal standard in such an anti-merger suit by creating, as the Supreme Court later interpreted it, a new affirmative defense where "anti-competitive effects . . . are clearly outweighed in the public interest by the probable effect of the transaction in meeting the convenience and needs of the community served." Subsequent decisions have defined this defense in a rigorous way that defendants have found extremely difficult to satisfy.[21]

Bank mergers can thus be seen to illustrate, unambiguously and by explicit statutory command, a regime of concurrent administrative and judicial jurisdiction:

(1) The Justice Department has the power to ask for a judicial antitrust veto of a merger; this power might or might not be exercised in any particular case.

(2) Administrative action is an affirmative precondition to consummation of the merger and thus must be sought by the parties. This would be easy to understand if the regulators were confined to appraising a merger's effects on solvency while the ordinary antitrust tribunals would adjudge anticompetitive effects. In fact, however, the banking agencies are required to consider competitive effects and to reject those mergers threatening competition unreasonably. Bear in mind, moreover, that these agencies (together with state agencies) control entry into the banking business. And this control over market structure is exclusive with the regulators; the antitrust court has no role to play there.

Why then should the antitrust court be given an independent role on those market structure questions arising in a merger approved by the regulators? Considerations of efficiency and orderliness in decision-making were given very little weight in the most recent Congressional enactment affecting this question. Everyone correctly perceived that the administrative agencies were more hospitable to bank mergers than the Justice Department and the courts. Accordingly, banking circles sought to minimize the role of the antitrust agencies. And others wanted to preserve that antitrust role.

---

[19] United States v. Philadelphia National Bank, 374 U. S. 321 (1963).

[20] Some courts think that there is a different test of legality under the Clayton and Sherman Acts. But there is no intrinsic distinction between the Clayton Act prohibition of mergers that "may tend substantially to lessen competition" and the Sherman Act prohibition of mergers that constitute an "[unreasonable] restraint of trade."

[21] United States v. Third National Bank in Nashville, 390 U. S. 171 (1968).

The fundamental truth is this: The decision to approve or disapprove a given transaction will often depend upon the perceptions and "institutional" outlook of the decision-maker in the light of what he considers most important. Thus, who decides will often determine what is decided. The regulator charged with encouraging competition may give competitive considerations less weight than the Justice Department and the courts.

☐ **The New York Stock Exchange.** The *Silver* case[22] involved these facts: Without a hearing or an explanation, the competing brokers comprising the NYSE discontinued their direct wires with plaintiff Silver, a nonmember broker, who sought an injunction and treble damages under the antitrust laws. Such a collective refusal to deal would be presumptively unlawful under the Sherman Act, but the NYSE acted pursuant to a rather clearly implied provision of its rules, and self-government by the NYSE was certainly assumed by Congress in enacting the Securities Exchange Act of 1934. Although Congress did not expressly prohibit or permit any particular NYSE rule, Sec. 19b of the Act authorized the SEC "to alter or supplement" NYSE rules "in respect of" 12 enumerated matters "and similar matters." (Sec. 19b9 gives the SEC jurisiction over NYSE rules concerning "the fixing of reasonable rates of commission.") The SEC had jurisdiction, which had not been exercised, to change the NYSE rule applied to Mr. Silver.

Although the Act did not explicitly exempt NYSE rules from the antitrust laws, the Supreme Court did not simply apply standard boycott doctrine: "The entire public policy of self-regulation beginning with the idea that the Exchange may set up barriers to membership, contemplates that the Exchange will engage in restraints of trade which might well be unreasonable absent sanction by the Securities Exchange Act." Nor did the Supreme Court permit the existence of unexercised Commission power over NYSE rules to immunize NYSE collective decision-making from the antitrust laws: "[T]he statutory scheme of that Act is not sufficiently pervasive to create a total exemption from the antitrust laws." Rather, the Court held the crucial question to be this: Was the NYSE rule in question necessary to achieve the purposes of the Securities Exchange Act? "[E]xchange self-regulation is to be regarded as justified in response to antitrust charges only to the extent necessary to protect the achievement of the aims of the [Act]."

In the application of that standard, the views of the expert body charged with responsibility in this area would seem relevant. Instead of referring the question to the Commission, however, the Court purported to distinguish Commission jurisdiction to alter NYSE rules for the future from jurisdiction to review a particular application to Mr. Silver. But there is not really a difference; the

---

[22] Silver v. New York Stock Exchange, 373 U. S. 341, 360–362 (1963).

Court, in effect, ruled as to the future when it held that no statutory objective prevented the NYSE from "telling a protesting nonmember why a rule is being invoked so as to harm him and allowing him to reply in explanation of his position. . . . The requirement of such a hearing will . . . [discourage] anti-competitive applications of exchange rules."

The Court did not instruct us as to how much the NYSE had to tell Silver, or what suspicions of unscrupulous conduct would be sufficient (indeed, whether anything less than clear proof of illegal conduct by Silver would justify cutting his wires), or whether such doubts, coupled with the peril of treble damages, might discourage the Exchange from acting at all against brokers who appear to be unscrupulous if not criminal.[23]

☐ **Summary.**

(1) Where an agency possesses an explicit statutory power to exempt private behavior from the antitrust laws, then behavior so exempted will indeed be immune from an antitrust suit. But such agency action will usually be subject to judicial review; and the reviewing judges will have to be satisfied that the agency followed proper procedures, that its decision was supported by evidence, that the agency's judgment was not unreasonable, and that the agency's judgment was authorized by and consistent with the governing regulatory statute. In interpreting the regulatory statute and in making these judgments, the judges are influenced by their understanding of national competitive policy as reflected in the antitrust laws.

(2) Where private parties have not sought an antitrust exemption from an agency empowered to give one, they will probably remain subject to some antitrust sanctions. The court may refer the problem to the regulatory agency in lieu of fashioning injunctive antitrust relief looking to the future. But criminal and damage antitrust remedies probably remain available through the courts, although the agency may be asked to find some relevant facts and even to make some relevant judgments.

(3) Similarly, the findings and judgments of the regulators may be relevant to a court's application of antitrust standards to regulated parties.

(4) Where the regulatory statute does not provide an express antitrust exemption for conduct approved by (or subject to) the regulatory agency, should an exemption be implied? Courts often ask, How pervasive is the regulatory scheme? And notwithstanding possible referrals to the agencies in the *Pan Am* or

---

[23] Kaplan v. Lehman Bros., 371 F. 2d 409 (7th Cir. 1967), was an investor suit against brokers who, through the NYSE, had agreed to fix prices, i.e., their minimum commission charges. The court did not refer the question to the Commission but held such price fixing authorized by Sec. 19b9, although Sec. 19b expressly authorizes the SEC to interfere rather than expressly empowering the Exchange to fix prices. Nor had the SEC in fact previously reviewed or controlled the NYSE minimum commission rule. The court did not decide that NYSE price fixing was necessary to achieve the Act's purposes. And that same court narrowed its Kaplan decision by later denying an automatic antitrust exemption for the NYSE rule forbidding rebates to nonmember brokers: Thill Securities Corp. v. NYSE, 433 F2.d 264 (7th Cir. 1970).

Far East situations, the most frequently repeated dogma is this: Exemptions from the antitrust laws are not lightly implied. Nevertheless, as in the *Silver* case, the Courts do recognize that the basic antitrust standard of "reasonableness" is sufficiently flexible to take account of the special circumstances of a regulated industry.

(5) Ultimately, peaceful coexistence between regulatory and antitrust regimes is possible because the federal courts interpret both the regulatory and antitrust statutes. For example, the NYSE minimum brokerage commission rule might come before the court as a review of SEC action approving or disapproving such a rule. Or the question might arise in an antitrust suit, with or without reference to the SEC by the court. In either case, the ultimate question is one of interpreting the Securities Exchange Act in the light of antitrust policy, and the federal judges will decide it.

5. State-regulated firms

■ The existence of state regulation affects antitrust application in two ways. First, private action addressed to the state authorities can, under some circumstances, constitute immune "political action." Second, where competition is restrained, not by private decisions but by the decisions of state legislative or administrative bodies, the restraint will often constitute "state action" to which the Sherman Act is not directly addressed.

☐ **Political action.** In the *Noerr* case,[24] the Supreme Court immunized a collective undertaking by competing railroads to obtain legislative and executive action from the state of Pennsylvania prejudicial to the railroads' truck competitors. Less than candid advertising was used, and anticompetitive motivation was clear. Nevertheless, any anticompetitive results were embodied in Pennsylvania legislative and executive decisions. The collective petitioning and lobbying was considered by the Court to be an inevitable, if not indispensable, incident of democratic government. The Court declined to distinguish between individual and collective "petitioning," perhaps because collective action is so often the essence of politics.[25]

It should be apparent, however, that the existence of regulation cannot immunize all private action directed to the regulatory agency. Suppose, for example, that a state regulatory agency were considering a possible regulation requiring all lenders to make certain disclosures or perhaps limiting certain creditor remedies. Suppose that the major lenders operating in that state made it clear that the proposed regulation would increase the cost of doing business and therefore lead to higher charges for borrowers. Suppose they discussed among themselves the precise amount of the prospective price increase, and either agreed on an *X*-percent response if the regulation were to be adopted or stated their collective conclusion that a specified price increment would "probably be required." Such collective price determination that is not indispensable to the political process should be held within the normal Sherman Act prohibition. Similarly unlawful would be an agreement among competitors to file false data with a regulatory commission in order to induce the commission to make a ruling harmful to their rivals.[26]

---

[24] Eastern Railroad Presidents Conference v. Noerr Motor Freight, Inc., 365 U. S. 127 (1961).

[25] In United Mine Workers v. Pennington, 381 U. S. 657 (1965), the Court similarly immunized an allegedly malicious conspiracy to persuade the Secretary of Labor to set minimum wage rates allegedly designed to harm the smaller coal producers. Private persons or groups seeking to persuade a government body to establish purchasing specifications—which would have the effect of excluding others—have been held outside the Noerr case's protection of political action addressed to government in its sovereign capacity and not as mere purchaser. See George R. Whitten, Jr., Inc. v. Paddock Pool Builders Inc., 424 F2d 25 (1st Cir. 1970); contra: United States v. Johns Manville Corp., 259 F. Supp. 440 (E. D. Penn. 1966).

The Supreme Court recently decided a case in which competing truckers alledgedly created a joint fund to oppose each and every application from any other trucker seeking a competitive operating certificate from either the Interstate Commerce Commission or the relevant California commission.[27] The result was to force every rival entrant to bear the delay and expense of lengthier and more complex administrative proceedings regardless of the merits of the application. Entry was thus made more difficult for outsiders and more costly for any applicant than for the established truckers, who split the cost of such proceedings. To be sure, the railroads were also trying to make life difficult for the truckers in the *Noerr* case, but there was also the offsetting social interest in keeping political channels open. And the railroads could reasonably claim that they were genuinely seeking political action; indeed, they got it. But the trucking group was not seeking administrative action in good faith. The decision to oppose every application regardless of the merits reflected an effort to impose litigation costs on applicants regardless of the administrative decision to grant or deny the license. The Court of Appeals for the Ninth Circuit refused to immunize this outrageous conduct from the antitrust laws, and the Supreme Court has affirmed that decision.

☐ **State action.** That the antitrust laws are addressed to action by private parties and not action by state legislatures or state administrative bodies was declared in the *Parker* case.[28] There, California's scheme for raising and fixing raisin prices had been challenged under the antitrust laws. To be sure, California itself set the price and marketing limitations, but it did so at the request and upon the consent of a majority of the growers. The effects were clear. California produced half the world's supply of raisins and most of that consumed in the United States. When California controlled the production, price, or marketing of raisins, interstate commerce was clearly and substantially affected. Thus, there were sufficient interstate effects to empower Congress to override state legislation pursuant to its supreme Constitutional power over interstate commerce. The question is, had Congress done so? To what extent do the antitrust laws express a federal policy sufficient to invalidate state laws establishing an anticompetitive regime that would be unlawful if effected solely by private persons? The answer to that question in any particular situation may depend upon (1) the strength and clarity of the federal policy, relative to (2) the strength and legitimacy of the state interest.

---

[26] Woods Exploration & Producing Co. v. Aluminum Co. of America, 438 F.2d 1286 (5th Cir. 1971).

[27] Trucking Unlimited v. California Motor Transport Co., 432 F.2d 755 (9th Cir. 1970), aff'd,—U. S.—(1972).

[28] Parker v. Brown, 317 U. S. 341 (1943).

Of course, the antitrust laws do not constitute the sole expression of federal policy with respect to competition. In some instances, federal law cedes paramount control to the states, as it does with respect to resale price maintenance and many aspects of the insurance or liquor business. In the *Parker* case itself, the Court did not ask whether the antitrust laws amounted to a preclusive exertion of federal power in favor of competition. It did not have to, because it found the California regulation consistent with national agricultural policy. The Agricultural Adjustment Act had authorized the Secretary of Agriculture to impose similar marketing restrictions. Moreover, Congress recognized the simultaneous existence of state regulation in the general statutory instruction to the Secretary to harmonize state and federal regulation. Thus, said the Supreme Court, the state regulation coincided with Congressional agricultural policy. It should be recognized, however, that such statutory recognitions of state regulatory interest are often vague in content and indefinite in scope.

The statutes governing FCC and FPC activities recognize the existence of state regulation of local power and telephone service. But how much does this fact illuminate the proper degree of federal toleration for state administrative decisions that might, for example, permit an electric utility to establish particular promotional rate structures discouraging consumers from installing alternate energy sources, or that permit a telephone company to prevent subscribers from installing auxiliary devices manufactured by others? If federal law recognizes the existence of some room for state utility regulation, how are the federal antitrust courts and state agencies to define the permitted scope of state control?

We cannot attempt to define the nature of a legitimate state interest without examining the history and application of the Constitution's interstate commerce clause. That is obviously a task well beyond this paper. But even without detailed examination of which objects are legitimate, two subordinate inquiries will resolve some cases. First, there should be no validation through inaction; there should be no antitrust immunity to the state-regulated firm whose conduct is not actually considered and approved by a state agency.[29] Second, a claimed objective can be assumed to be legitimate for purposes of the analysis, which can then ask whether the regulatory decision restricts competition more than necessary to achieve the assumed objective. In the *Stiffel* case,[30] the Supreme Court held that Illinois could not prevent Sears from copying Sitffel's unpatented and unpatentable lamp, for that would interfere too much with the federal policy in favor of competition apart from the circumscribed area of federal patent law. Now if the copied lamp were identified in the consumer's mind with Stiffel, copies might appear to be a Stiffel product and thus might mislead buyers. Preventing copying would avoid such deception, but so would labeling, which is a far lesser impediment to competition.

---

[29] The opposite conclusion was reached by Washington Gas Light Co. v. Virginia Electric and Power Co., 438 F.2d 248 (4th Cir. 1971), where, as an alternative holding, the Fourth Circuit Court of Appeals exempted from the antitrust laws a possible "tie-in" promotional arrangement as protected by the Parker case, even though the state agency had not considered the practice. In a question-begging formulation, that court assumed (1) that surveillance of utility practices was the duty of the state agency, (2) that the agency should be presumed to have performed its duty, and therefore (3) that failure of the state agency to object to practices it presumably knew about was tantamount to state approval. This reasoning is unsound and has already been gently questioned by the Court of Appeals for the District of Columbia. See Hecht v. Pro-Football, Inc., 1971 Trade Cases, para. 73, 559 at n.17 (D. C. Cir.).

Thus, without challenging the state's right to prevent deception, a federal court might ask whether the restriction of competition is both related to that objective and essential to it. Sometimes the state legislative standard will be appropriate, as where it restricts the practice of professions to competent persons, but the administration of that standard may not be, as where the examinations are too "hard" or not particularly relevant to competence. Although the courts have always hesitated to second-guess the application of such standards, they have increasingly probed beneath the surface of state administrative and legislative judgments in the civil rights area. To be sure, federal concern with competition is probably thought to stand lower in the hierarchy of federal values than the protection of personal liberty. But once the courts have recognized that the assumed difficulties in probing behind the appearance of state action are not insurmountable, earlier compunctions against close review of state economic action may be weakened.

The critical difficulty here lies in framing the proper standard by which federal courts are to review state action. Before the New Deal, the courts were readier to invalidate state economic regulation on "substantive due process" grounds. In those days, the courts would judge whether the state interference with private economic freedom was insufficiently justified and was therefore likely to "deprive any person of . . . liberty or property without due process of law." Since 1937, however, the courts have tended to take a rather hands-off attitude toward state economic legislation. The interesting question for the future is this: When and to what extent will the federal courts, and the Supreme Court in particular, come to see that state regulation may be offensive, not in terms of the discredited laissez-faire notions of substantive due process, but in terms of depriving the national economy of the competitive presumptions established by the federal antitrust laws. Of course, hard questions would be opened, and the courts may choose to let sleeping dogs lie. They have enough troubles in other areas, and the present hodgepodge does work, more or less. It is clear, however, that existing decisions do not begin to scratch the surface of the relationship between federal antitrust and state regulatory law. I shall await the enterprising plaintiff who seeks direct federal review (not an antitrust suit) of state denial of a taxicab license, not on due process grounds, but on grounds that the state taxi cartel affects interstate commerce significantly without being necessary to assure safe vehicles or competent and identifiable drivers.

---

[30] Sears Roebuck & Co. v. Stiffel Co., 376 U. S. 225 (1964).

6. Conclusion

■ The summary points are few. First, much more thought has to be given to relations between state laws limiting interstate competition and the general federal policy, as expressed in the antitrust laws, favoring competition. Second, it appears that procedure often determines substance. Much regulation is unnecessary, carried too far, or implemented with more anticompetitive incidents than warranted by the basic rationale of regulation. Although the judges often perceive this, the legacy of the New Deal era is that courts are supposed to leave economic matters to legislatures and administrators. This legacy will often lead the judges to act in ways contrary to their perceptions of the economic issues. They seek harmonious coexistence between competitive and anticompetitive regimes in a variety of doctrines about state action, administrative expertness, or primary jurisdiction. The upshot is a measure of chaos and, frequently, who decides in the first instance will often determine what is decided.

The application of finance theory to public utility rate cases

Stewart C. Myers

Associate Professor of Finance
Sloan School of Management
Massachusetts Institute of Technology

The purpose of this paper is to formulate the implications of finance theory for rate of return regulation. A variety of problems in finance and the law and economics of regulation are reviewed. Also, a regulatory procedure based on finance theory is proposed for practical use.

Finance theory suggests that the "comparable earnings" standard for rate of return regulation ought to be based on utilities' costs of capital. The cost of capital is difficult to measure, since it is defined in terms of investors' expectations. But plausible estimates can be obtained for utilities. The following principle is proposed for use of these estimates: Regulation should assure that the average expected rate of return on desired new utility investment is equal to the cost of capital. This is a definition of "fair return" based on the theory of competitive equilibrium. The principle is consistent with the comparable earnings standard. Only the most obvious, "straightforward" approach to implementing this principle is examined in detail; but this approach is practical and logically sound. It is particularly attractive when combined with conscious use of regulatory lag as an instrument of regulation. There are some difficulties, however, when such a lag is combined with the usual regulatory practice of basing the allowed rate of return on embedded rather than current debt costs.

The problem of determining the appropriate rate base is also discussed. Regulation based on the book value rate base will not generally lead to efficient price, output, or investment decisions. A "competitive market value" rate base would be better from the standpoint of efficiency, since it would lead to long-run marginal cost pricing. However, long-run marginal cost pricing is not generally consistent with the principle that utilities ought to be able to expect to earn their cost of capital on new investment.

1. Introduction

■ There is little argument in practical circles about the broad purpose of regulating public utilities' rates of return. The accepted legal

---

Stewart C. Myers received the A.B. degree from Williams College (1962) and the M.B.A. and Ph.D. degrees from Stanford (1964 and 1967, respectively). He is associate editor of the *Journal of Finance* and the *Journal of Financial and Quantitative Analysis* and has authored or coauthored several articles and a book in the field of finance.

An earlier draft of this paper was presented at the conference on "Problems of Regulation and Public Utilities" held at Dartmouth College in September, 1970. The author is grateful to seminar participants for helpful criticism. Particular thanks are also due to Willard Carleton, David Schwartz, Haskell Wald, and the editors of this Journal for their detailed comments. They are not responsible for any mistakes or misguided opinions. Parts of this paper were also included in "AT&T Cost of Capital Study," presented as testimony by Myers as a staff witness in the 1971 interstate telephone rate case [53].

principle is that "the return to the equity owner should be commensurate with returns on investments in other enterprises having corresponding risks," and sufficient to "attract capital" and maintain credit worthiness.[1]

Finance (the study of investments, capital markets, and financial management) seems directly relevant to the regulatory agencies' attempt to determine the rate of return allowed by law. Although relevant, the applications of finance theory in the regulatory field are not easily seen or understood.

The difficulties involved in transferring finance theory to regulatory problems are attested to by the sharp controversies generated by economists who have ventured into rate proceedings. Disagreements have arisen not only between economists and the usual participants in these proceedings but also among the economists themselves. Thus the problem is not simply *explaining* the theory to the regulators; it has not been clear how the theory should be applied.

The purpose of this paper is to formulate the implications of finance theory for public utility regulation. It is an attempt to apply theory to practical affairs. Thus, detailed attention is given not only to the theory per se but to current regulatory procedures, to alternative procedures, and to the controversies associated with recent attempts to apply finance theory in the regulatory arena.

A few caveats are necessary at the start:

(1) The paper's title means just what it says. This is not a general treatment of finance theory or of public utility rate regulation, but of the application of the former in the latter field. The various aspects of finance and regulation are pursued only as far as is necessary for this limited purpose. For example, I treat the practical problems of estimating utilities' costs of capital only superficially, concentrating instead on how an estimated cost of capital should be used. The problem of estimation is at least as hard and important as that of application, but it is substantially the same problem faced by nonregulated firms. Reviewing it in detail would mean reviewing much of finance theory in detail—a task I cannot attempt here.

(2) I will concentrate on one possible method of regulation, namely control of the overall rate of return earned. For the most part, I am taking the existing legal and procedural framework as given. I do not mean to imply that this framework is necessarily the best one.

(3) The paper does not necessarily describe how regulators actually behave, but how they should behave if the legal and procedural framework is taken at face value and the implications of finance and economic theory are recognized. One suspects that various nonfinancial considerations also affect regulatory behavior, and that regulators respond to objectives that are not evident in the legal and procedural framework.[2]

In short, I do not propose to present an optimal regulatory strategy or a complete positive theory of regulatory behavior. The

---

[1] This is from the Supreme Court Decision on Federal Power Commission et al. v. Hope Natural Gas Company, 320 U. S. 591 (1949) at 603.

[2] For example, see Posner [38] and Stigler [48] for substantially different views of the role of regulation.

objective is to provide an analysis which can be used to improve existing procedures. This is a limited, but important, objective. Existing procedures not only involve Big Money, but they apparently influence the allocation of resources to, and within, a substantial segment of the economy.

The paper begins with a brief review of the mechanics of rate of return regulation and describes the traditional interpretation of the accepted legal standards for a "fair" rate of return. I argue that the traditional interpretation has serious deficiencies, both in logic and in application. A market-based cost of capital is suggested as an alternative basis for establishing a "fair" return.

The middle portion of the paper (Sections 3 through 5) discusses the definition and measurement of the cost of capital, under the simplifying assumption that the firm is all-equity financed. A number of controversies involving definition and measurement for a regulated firm are discussed.

Measuring the cost of capital is one question, using it another. Sections 6 and 7 summarize the aims of regulation and show that these aims cannot be achieved solely by regulating overall rate of return. However, a somewhat limited objective is logical and feasible; that is, regulation should insure that the expected rate of return on desired new investment is equal to the cost of capital. I present an approach to regulation which is consistent with this principle and discuss the real and alleged problems in implementing it.

The penultimate section drops the assumption of all-equity financing, briefly discusses the problem of determining optimal capital structure, and shows how the approach developed in Sections 6 and 7 can be applied to utilities with complicated capital structures.

The last section includes a brief summary of major conclusions.

2. Review of rate regulation

■ **Regulatory procedures.** A utility's prices are set so that the utility covers its costs, including taxes and depreciation, plus a certain return on investment. The return on investment is obtained by multiplying the "rate base" by an allowed rate of return. The rate base is essentially the book value of the utility's capital investment.[3]

For example, suppose Utility X has a rate base of $100 million. It is producing one billion widgets per year. The allowed rate of return is set at 10 percent overall. Costs are $50 million per year, including depreciation and all taxes.[4] Then the average price of widgets is set as follows:[5]

$$\text{Price per widget} = \frac{\text{Revenue requirements}}{\text{No. of widgets}} = \frac{50{,}000{,}000 + 0.10(100{,}000{,}000)}{10^9} = \$0.06\,.$$

Thus widgets are sold for $0.06 until the next regulatory proceeding. This procedure allows the actual return earned to be more or less than

---

[3] Sometimes "fair value" rate bases are used. They are discussed later in this paper.

[4] That is, costs would be adjusted to cover income taxes associated with a 10-percent return, as well as property taxes, social security contributions, etc.

[5] This paper is concerned with the *level*, not the structure, of utility rates.

TABLE 1

PACIFIC GAS TRANSMISSION CO.–ALLOWED OVERALL RATE OF RETURN

CAPITAL SOURCE	PERCENT OF CAPITALIZATION	RATE OF RETURN IN PERCENT	WEIGHTED RATE
DEBT	79.66	6.13	4.88
EQUITY	20.34	11.64	2.37
TOTAL	100.00		7.25

SOURCE FPC OPINION NO. 569 [58], P. 15.

10 percent, depending on realized cost and revenues during the period of "regulatory lag" between proceedings.

The allowable rate of return is computed in the same fashion as a weighted average cost of capital, but with important differences. Table 1 shows the computation of the overall rate of return allowed Pacific Gas Transmission Company in a 1970 Federal Power Commission decision. The procedure for computing the overall return, 7.25 percent, is clear from the table, but the table does not show where such numbers typically come from.

(1) The figures listed under "percent of capitalization" refer to the respective percentages of debt and equity listed on the company's books at the time of the rate proceeding. They are book, not market, values. Neither do they necessarily refer to the proportions of debt and equity to be employed in future financing.

(2) The 6.13-percent figure is the "embedded" debt cost—that is, total interest payments divided by the book value of outstanding debt. The company's current borrowing rate can be much higher or lower than the "embedded" debt cost shown.

(3) The percentage cost of equity is a figure arrived at by judgment. I discuss the possible bases for this judgment below.

Although there are opportunities for disagreement at each of these three steps, most argument is centered on the return to equity, so I will start there.

☐ **The legal standard for setting the return to equity.** The governing principle is the Supreme Court's statement in the Hope decision:

> The return to the equity owner should be commensurate with returns on investments in other enterprises having corresponding risks. That return, moreover, should be sufficient to assure confidence in the financial integrity of the enterprise, so as to maintain its credit and to attract capital.[6]

The first sentence establishes a standard of "comparable earnings," the second, a standard of "capital attraction." The broad language of the Hope decision allows a variety of specific interpretations of these standards.

It is best to start with the comparable earnings standard. In practice, it has been applied in two ways. The traditional and most widely accepted approach[7] defines "returns" as the book rates of return of other firms.[8] Those who advocate the traditional approach read the Hope decision in a particular way, namely:

---

[6] Note 1 *supra.*

[7] The most complete exposition and defense of this method is Leventhal [22].

> The return to the equity owner should be commensurate with [*recent book*] returns on [*past*] investments [*made by*] other enterprises having corresponding risks.

This approach rests on a special notion of opportunity cost—in this context, that a utility should be allowed to earn what it *would have* earned had its capital been invested in other firms of comparable risk.[9]

The alternative suggested by finance theory is to define "commensurate return" as the rate of return investors anticipate when they purchase equity shares of comparable risk. This is a *market* rate of return, defined in terms of anticipated dividends and capital gains relative to stock prices.

☐ **Drawbacks of the traditional interpretation.** The traditional interpretation has clear deficiencies. First, the method does not rest easily on the concept of opportunity cost. Opportunity cost is a marginal and forward-looking concept. Thus, the opportunity cost of investing in a particular asset is usually defined as the anticipated rate of return on an incremental investment in the best alternative. However, observed book rates of return are average returns on past investments.

Granted, average and marginal returns are equal at long-run equilibrium in perfectly competitive markets; but this is in an *ex ante* sense. No one argues that perfect competition requires equality of rates of return after the fact in an uncertain world. In any case, usually no attempt is made in regulatory proceedings to see whether the data examined really are marginal rates of return and whether they stem from perfectly competitive situations.

Second, the traditional interpretation of comparable earnings ignores capital markets. This is serious because the Supreme Court specifies that the variable of interest is "the return to the equity owner." The shareholder is not directly interested in the ratio of book earnings to the book value of a company he invests in. He looks at anticipated dividends and capital gains relative to the stock price he has to pay. Thus, it is more relevant to interpret the opportunity cost of capital as the return on securities with risks similar to the stock of the utility in question.

A third objection is that it is difficult, in practice, to find a suitable class of firms with corresponding risks. Suppose you are looking for a company with risk commensurate with Utility X. The likeliest candidates are other utilities, whose reported book rates of return reflect past regulatory decisions as well as competitive forces. Basing current regulatory decisions partly on past decisions leads to a dangerously arbitrary standard. One is forced to look elsewhere, to unregulated firms. But such firms presumably are riskier than utilities. Moreover, there is no clear theory about how risk should be related to differences in book rates of return (or even how risk should be defined if the book rate of return is the variable of interest.) In contrast, the relationship of risk and return in capital markets is better understood. This is discussed later in the paper.

---

[8] I will use the term "book" to refer to data based on income statements and balance sheets. Thus, the book rate of return to equity is simply the ratio of reported income to net worth as shown on the firm's balance sheet.

[9] The reader may wish to judge for himself whether this characterization of the "traditional interpretation" is accurate. I would suggest starting with two intelligently done examples of this approach. See testimony of Solomon [60] and Friend [37]. Friend concludes: "If regulation is attempting to duplicate competitive results, the rate of return permitted on AT&T equity should be of the same general order of magnitude as on industrials and electric utilities with the same [risk] characteristics. This would require a return in the neighborhood of 12% to be consistent with [book] returns achieved in the last decade by electric companies and even somewhat higher to be consistent with the return achieved by comparable industrials" (p. 17).

The final objection is that accounting rates of return are subject to serious measurement errors and biases, particularly when comparisons are made between firms in different industries. This is also discussed later in the paper.

To summarize, the difficulties with the traditional interpretation of the comparable earnings standard are at very least sufficient to justify examination of alternatives.

3. Finance theory and the cost of capital concept

■ If a utility's allowed rate of return is to be "sufficient . . . to attract capital" and "commensurate with returns on investments in other enterprises having corresponding risks," then it has something in common with the cost of capital as that concept is used in finance. But "the cost of capital" is one of those phrases that can mean a dozen things. Thus, it will be helpful to review the concept briefly.

☐ **Definitions and assumptions.** To simplify matters, we will concentrate for the moment on a firm that is all-equity financed and can ignore market imperfections such as transaction costs and taxes. The logic in developing a cost of capital (i.e., minimum acceptable expected rate of return, or "hurdle rate") for such a firm's investments goes as follows:

(1) The firm is one of a class with similar risk characteristics—call this class j.

(2) At any point in time there is a unique expected rate of return prevailing in capital markets for this degree of risk—call it R_j.

(3) The share price of the firm in question will adjust so that it offers an expected rate of return R_j to investors.

(4) This rate, the shareholders' opportunity cost, should be the minimum acceptable expected rate of return on new investment, assuming the projects under consideration have risk characteristics similar to currently held assets. Otherwise, the firm's shareholders' wealth will not be maximized.

This sequence of logic defines the appropriate discount rate for projects which do not change the firm's risk characteristics. The basic problem is one of estimating the rates prevailing in the market.

The following equilibrium condition will be assumed to define R_j:

$$P_{tj} = \frac{D_{t+1,j} + P_{t+1,j}}{1 + R_j}, \qquad (1)$$

where

P_{tj} = ex-dividend price at the end of period t;
$D_{t+1,j}$ = dividends[10] expected to be received in $t + 1$, and
$P_{t+1,j}$ = expected ex-dividend price at the end of $t + 1$.

It is not literally true that everyone has the same expectations of future returns. However, for purposes of analysis I assume that it is permissible to speak of "the market's" expectations.

Also, note that if R_j is assumed constant over time,[11] equation (1) implies

$$P_{0j} = \sum_{t=1}^{\infty} \frac{D_{tj}}{(1 + R_j)^t}. \qquad (2)$$

☐ **Additional comments.**

Risk classes

The phrase "risk class" does not strictly imply that risk can be measured in one dimension. Moreover, it is conceivable that equilibrium expected rates of return on securities depend not only on risk but on still other factors, although this has not been established.

Nevertheless, the concept of opportunity cost (relative to an equivalent class of securities) is not made invalid by the fact that it may have to be defined as a function of several variables.[12] It is true, of course, that the more complicated the function, the more difficult is use of the concept.

Evaluating investments in differing risk classes

Most of the analysis in this paper assumes that the firm's risk class is given. However, what happens if it is not?

Suppose the firm acquires or disposes of an asset or liability, and that the transaction changes the firm's risk class from j to w. Assume that the asset or liability considered separately is in class k. Then in perfect markets,

$$P_{wt} = \bar{P}_t + \Delta P_t = \frac{\bar{D}_{t+1} + \bar{P}_{t+1}}{1 + R_j} + \frac{\Delta D_{t+1} + \Delta P_{t+1}}{1 + R_k}, \qquad (3)$$

where $^{-}$ indicates initial values and Δ, changes due to adoption of the project in class k.[13] Alternatively,

---

[10] "Dividends" must be broadly interpreted to include all cash flows from the firm to holders of the share in question. Conceivably, D_t might include return of capital or even direct repurchase of shares by the firm.

[11] This assumption avoids consideration of the term "structure of interest rates." It also assumes that the perceived risk of successive future dividends increases at a constant rate as a function of t. See Robichek and Myers [40].

[12] For example, the different tax treatment of capital gains vs. dividends implies, *ceteris paribus*, that equilibrium expected rates of return should be positively correlated with dividend yield. If so, this would add another dimension necessary to define a class of equivalent securities. However, it is not clear whether there is any systematic empirical relationship between dividend yield or payout and equilibrium rates of return. See, for example, Friend and Puckett [15], Miller and Modigliani [30], especially p. 370, and Black and Scholes [6]. Others, notably Gordon, have argued that equilibrium rates of return are negatively related to dividend payout. See [16], for example.

$$\Delta P_t = \sum_{t=1}^{\infty} \frac{\Delta D_t}{(1 + R_k)^t}, \tag{4}$$

where ΔD_t is the expected incremental cash flow from adoption of the project. Thus, if the asset or liability under consideration has risk characteristics like securities in class k, then the appropriate opportunity cost is R_k, the equilibrium rate of return offered by class k securities.

4. Estimating the cost of capital

■ This paper does not go very deeply into the problem of estimating investors' opportunity costs. However, a brief review will be helpful in two ways. First, it will show the kinds of evidence likely to be relied on in a practical context. Second, it will show some of the implications of the view of security valuation presented in Section 3 above.

The basic proposition underlying the cost of capital concept is that at any point in time securities are so priced that all securities of equivalent risk (i.e., all securities in a "risk class") offer the same expected rate of return. For a given utility the basic problem is to determine the expected rate of return for the class in which the stock falls.

There is no mechanical way to do this. Measurement of expectations is intrinsically difficult. But there are several types of evidence that should be examined before the ultimate judgment is made.

1. Interest rates

Interest rates on corporate bonds and other debt instruments can be readily observed to provide a floor for the estimate. Changes in the basic level of interest rates normally correspond in direction to changes in the cost of equity capital.

2. Ex post rates of return to investors

Averaging of *ex post* rates of return (or better, of *ex post* risk premiums, since interest rates vary over time) give some indication of the relevant range in which expectations lie. These averages are most helpful to the extent that they cover a long period of time and many stocks.[14] One cannot very well rely on five years of history for the utility in question as a guide to investors' expectations for the future.

3. DCF formulas

Examination of interest rates and past rates of return indicates a range for expected rates of return on common stocks. But these measures give insufficient bases for estimating a particular firm's cost of equity capital.

---

[13] The theorem is proved in a more general form in Myers [36]. There are, of course, degrees of perfection, and there is a residual disagreement about how perfect markets have to be in order for equations (3) and (4) to hold. See Lintner [23], p. 108, and on the other hand, Hamada [19].

[14] There is plenty of evidence available. See, for instance, Fisher and Lorie [13].

One approach is the so-called DCF or discounted cash flow method.[15] Its basic premise is equation (2)—i.e., that stock price is the present value, discounted at R, of the stream of expected dividends.[16] The idea is to infer R from the observed price P_0 and an estimate of what investors expect in the way of future dividends.

In practice a number of simplifications of equation (2) are employed in using the DCF method. Suppose, for example, that the dividend stream is expected to grow indefinitely at some rate g which is less than R. Then equation (1) can be simplified to

$$P_0 = \frac{D_1}{R - g}, \tag{5}$$

$$R = \frac{D_1}{P_0} + g. \tag{5a}$$

For utilities, for which a constant, moderate long-term trend in earnings and dividends is often identifiable, equation (5a) can be a reasonable rule of thumb for estimating R. A danger is that temporary growth trends are apt to be mechanically projected "to infinity." Likewise, it is tempting to assume without checking that expected future growth is constant. Fortunately, there is nothing in equation (2) that requires a single, perpetual growth rate. One can easily assume that different growth rates are anticipated for different future periods.[17]

In general, the DCF model—either equation (5a) or some more complicated variant of equation (2)—has to be fit to the case at hand. The point of the analysis is to answer the question, What would a rational unbiased investor expect from a long-term investment in this stock at the prevailing price? This rate of return is taken to be R on the assumption that the prevailing price is based on the opportunity cost of investment in equivalent-risk securities.

4. *Earnings-price ratios*

Earnings-price ratios can be used to measure the cost of equity capital in some cases. The formulas

$$P_0 = \frac{\mathrm{EPS}_1}{R}, \tag{6}$$

$$R = \frac{\mathrm{EPS}_1}{P_0} \tag{6a}$$

are actually special cases of equation (2) if certain assumptions hold.

Suppose that earnings per share (EPS) in any one year are "normal" long-run earnings of the firm's business and that all earnings are paid out as dividends. Then equations (6) and (6a) are simply equations (5) and (5a) but with $g=0$.

---

[15] For examples of the use of the method, see testimony of Brigham [55], Kosh [52], Myers [53, 61], and Roseman [57].

[16] The subscript j has been dropped because the firm's risk class is taken as given.

[17] In [61] I used a simple structural model of the firm to project the dividend stream under different assumptions about the short-term growth, long-term growth, and year-by-year book profitability. The dividend stream and the final estimate of the cost of equity capital were based on these simulations.

Thus it is said that EPS_1/P_0 measures the cost of equity capital for "no-growth" firms. This is possibly misleading, however. Suppose a firm which falls initially into the no-growth category instead reinvests a portion of its earnings in projects which have on the average a present value of exactly zero. Then announcement of these projects makes the firm no more nor less attractive to investors, even though the firm will expand because of the reinvested earnings. The firm's current stock price will not change. Therefore, R is still correctly measured by EPS_1/P_0.

If the projects are on the average more than marginally desirable, however, the price will rise, earnings per share will remain constant, and thus the earnings-price ratio will underestimate R.

Note that equation (2) can be written

$$P_0 = \frac{EPS_1}{R} + \sum_{t=1}^{\infty} \frac{D_t - EPS_1}{(1+R)^t}. \tag{7}$$

The second term can be interpreted as the net present value of future growth opportunities. Equation (6) follows from equation (7) only if the second term equals zero—i.e., if the firm's future investments yield exactly R on the average.

We see that growth in itself does not invalidate equations (6) and (6a). What does invalidate the formulas is growth that is more or less than minimally profitable.

This result has an interesting practical implication. Suppose it is argued that a utility's earnings-price ratio underestimates R. Then it must also be argued that investors expect the utility to earn more than the cost of equity capital on its future investments.[18]

☐ **Measuring risk.** It is clearly better to estimate R from data from a sample of equivalent-risk companies, rather than from data pertaining to the utility in question. But this requires an operational definition of "equivalent risk."

There is no consensus as to how risk should be defined and measured. Nevertheless, it is possible to obtain, from historical data, one or two statistics that are widely used as proxies.

The risk measures most often used stem from Markowitz's formulation of the individual's portfolio selection problem.[19] The investor is assumed to balance R_p, the expected one-period return on his portfolio, against σ_p^2, the variance of R_p. Let x_i equal the proportion of his investment allocated to security i, one of N candidates. Let σ_{ik} be the covariance of R_i and R_k and let $\sigma_i^2 = \sigma_{ii}$. Then

$$R_p = \sum_{i=1}^{N} x_i R_i, \tag{8}$$

---

[18] This assumes that EPS_1 equals expected average earnings from assets held in $t = 1$. EPS_1 may differ from earnings as actually reported.

[19] See Markowitz [28]. For more extended treatment of the concept and its applications, see Markowitz [29] and Sharpe [45].

$$\sigma_p^2 = \sum_{i=1}^{N} \sum_{k=1}^{N} x_i x_k \sigma_{ik} \tag{9}$$

with $\sum x_i = 1$.

Consider the special case in which the portfolio is limited to one security. For such an undiversified investor the relevant risk of security i is simply σ_i^2. Thus, if it is believed that demand for utility z's shares stems predominantly from undiversified investors—the proverbial widows or orphans who own stock in at most a few firms—then an estimate of σ_z^2 from past data should be a reasonable risk proxy, and a class of equivalent risk securities could be defined by $\sigma_i^2 \cong \sigma_z^2$.

However, it is hard to believe that the special case of widows and orphans is dominant. In general, the risk of security i is its marginal contribution to σ_p^2. This is

$$\frac{\delta \sigma_p^2}{\delta x_i} = 2x_i\sigma_i^2 + 2 \sum_{k \neq i} x_k \sigma_{ik}$$

$$= 2 \sum_{k=1}^{N} x_k \sigma_{ik} = 2\sigma_{ip} \, . \tag{10}$$

Thus security i's risk is proportional to σ_{ip}, the covariance of R_i and R_p.

As the number of securities in a portfolio increases, R_p becomes more and more closely correlated with R_M, the return on the "market portfolio" composed of all securities. (In fact, a high correlation of R_p and R_M has been found for randomly selected portfolios consisting of as few as ten stocks.)[20] This suggests σ_{iM} as an indicator of the "systematic" or "undiversifiable" risk of security i—the risk relevant to a diversified investor. σ_{iM} is in turn proportional to the coefficient β_i in the linear regression equation

$$R_i = \alpha_i + \beta_i R_M \, , \tag{11}$$

since β_i is given by σ_{iM}/σ_M^2.

Now, β_i can be estimated from past data. Thus it is a natural proxy for the effective risk of security to the well-diversified investor.

There is disagreement about the relative importance of β_i and σ_i^2 as determinants of R_i, but one and/or the other capture most of what most economists understand as risk.

☐ **The capital asset pricing model.** The risk measure β_i was just "derived" on a pragmatic basis, in order to show that it is *one* reasonable risk measure even in the absence of a formal theory of how risk affects security prices. Such justification is available, however, in the so-called capital asset pricing model.[21] Suppose we assume:

(1) That investors have identical assessments of securities' expected returns and risk characteristics,

(2) That the Markowitz model describes their portfolio choice, and

(3) That they can borrow or lend at a given risk-free rate, R_F.

---

[20] See Evans and Archer [11].

[21] The model is due to Sharpe [44], Lintner [24, 25], and Mossin [35].

Then at equilibrium in perfect markets,

$$R_i - R_F = \beta_i(R_M - R_F)\,,$$
$$\beta_i = \frac{\sigma_{iM}}{\sigma_M{}^2}\,. \tag{12}$$

It is not clear that equation (12) provides a *complete* empirical explanation of asset valuation.[22] But it does lend additional support to the use of β_i as a risk measure.[23]

Note also that if equation (12) is accepted it provides a vehicle for estimating R_i. R_F is approximately observable—a Treasury Bill rate, or a prime commercial paper rate, is customarily used. And presumably R_M, or $R_M - R_F$, will be easier to estimate from historical data than R_i or $R_i - R_F$. However, this approach has not yet been used in a regulatory proceeding.

☐ **Further comments on estimation in practice.** It is not my intention to go very deeply into practical problems of estimation. Nevertheless, a few comments may help to put the material just presented in better perspective. The problem is assumed to be estimating the opportunity cost of stockholders in Utility X.

It would be nice if investors' expectations were readily observable. If they were, they would surely be reported in the financial pages; and estimating R for Utility X would be a matter of looking up the currently projected dividends and capital gains, observing X's price, and calculating a rate of return. This would be taken as R_x on the assumption that X's stock is accurately priced relative to alternative equivalent risk investments. A sample of one firm (X) would be sufficient.

This is a never-never land. Suppose, however, we start with a sample of one firm, then postulate that a utility's future growth is relatively stable and predictable, and that it is therefore reasonable to use past trends as proxies for investors' expectations.

1. DCF estimates

Take the case of AT&T.[24] In March 1971 its annual dividend was \$2.60, the share price about \$49, and the dividend yield 0.053. During the 1960s the growth trend in its earnings per share was 4.6 percent. The trend in dividends per share for the same period was 4.5 percent. Suppose equation (5a) applies. If investors expected continued growth at about 4.5 percent per year, then their expected rate of return must have been $0.053 + 0.045 = 0.098$, that is, about 10 percent.

But it is immediately clear that the future need not be like the past. For example, AT&T's total assets may grow at a different rate,

---

[22] See Friend and Blume [14] and the several empirical studies in Jensen [20].

[23] Equation (12) also suggests that $R_i - R_F$ and $R_M - R_F$ should be substituted for R_i and R_M in estimating equation (11). Even if equation (12) does not hold exactly, security returns at any point in time depend on R_F, and a more reliable estimate of β_i can be obtained if fluctuations in R_F over time are adjusted for.

[24] The numerical examples following are drawn from my recent testimony [53]. I am not trying to summarize that testimony, nor am I implying that what was done there is necessarily appropriate for other cases. I am using that testimony here as a convenient source of numerical examples.

TABLE 2

COST OF EQUITY CAPITAL ESTIMATES BASED ON SIMULATION MODEL

AVERAGE BOOK RATE OF RETURN ON EQUITY INVESTMENT	LONG-TERM GROWTH RATE OF BELL SYSTEM ASSETS			
	0.07	0.08	0.09	0.10
0.100	0.097	0.098	0.098	a
0.105	0.101	0.102	0.103	0.104
0.110	0.106	0.107	0.108	0.109
0.115	0.110	0.111	0.112	0.113
0.120	0.115	0.116	0.117	0.118
0.125	0.119	0.120	0.121	0.122

[a] NO MEANINGFUL FIGURE EXISTS FOR THIS CASE

SOURCE SIMULATION RUNS, DESCRIBED IN [53], APPENDIX B.

and it may be more or less profitable. Since investors took such facts into account in their assessments of expected future earnings and dividends, equation (5a) may not apply.

2. *Simulations*

The obvious next step is to explore the consequences of alternative assumptions about the future performance of AT&T.

Table 2 shows the long-run rate of return from investments in AT&T stock at $49, under various assumptions about book rate of return on equity (ROI) and long-term asset growth (g_A).[25]

Table 2 should be read in the following way: If investors expect $g_A = 0.08$ and ROI = 0.10, then the cost of capital is 0.098, roughly the same as estimated previously via equation (5a). However, if investors expect $g_A = 0.08$ and ROI = 0.115, then the cost of equity capital must be 0.111; otherwise AT&T's stock would not sell for $49.

Thus, to the extent that it is possible to establish a reasonable range for investors' expectations of asset growth and book profitability, it is possible to specify a range for the cost of capital. For example, if there is no evidence that could justify an expectation of ROI > 0.12 or $g_A > 0.09$, then R must be less than 0.117.

Note that the cost of equity capital estimate increases with the assumed growth in assets. The higher g_A, the greater the present value of growth opportunities (ROI $> R$ in all instances) and the greater the cost of capital needed to explain the observed share price. However, the cost of equity capital estimate is much more sensitive to ROI than g_A, which reflects an interesting problem peculiar to regulated firms. The range of possible variation in ROI is wide partly because of uncertainty about the behavior of AT&T's various regulators. Granted, some of the more extreme values in Table 2 might be rejected as estimates of investors' *expectations*, but the uncertainty persists. This is one more reason why testimony in a regulatory proceeding cannot rest on a sample of one firm. The regulatory process introduces an element of uncertainty, which makes it difficult to assess investors' expectations, and thereby makes it difficult to measure the cost of capital. It is obviously necessary to broaden the sample.

---

[25] The table is based on a simulation which is described in detail in [53], Appendix B. The simulation was necessary because of the complexity of the relationship between g_A, ROI, and the projection of dividends and earnings per share. The major complicating element was the necessity for periodic stock issues to finance asset growth at g_A.

3. *Risk classes*

However, broadening the sample requires specification of a risk-equivalent class of stocks. Suppose we make use of the risk proxies described above. Figure 1 plots *ex post* return versus σ and β, respectively, for AT&T and Moody's 24 utilities, a sample of large, well-established electrics. The points shown were calculated from monthly rate of return data covering the 1960–1969 period.[26]

The figure shows that, compared to these utilities, AT&T was a relatively safe investment for the undiversified investor. For a well-diversified investor AT&T's risk was about the same.

Suppose that the 24 electrics are accepted as an "equivalent risk class." The logical next step is to estimate the cost of equity capital for the utilities. There are several ways in which this could be done. We might observe the average dividend yield of the 24 utilities (0.054) and the average of their 1960–1969 trends in earnings per share (0.06). Then using equation (5a),

$$R = \frac{D_1}{P_0} + g$$

$$= 0.054 + 0.06 = 0.114\,.$$

4. *The role of judgment*

One can go on to consider other companies, other measures, and other time periods. The only solid generalization is that, at the present state of the art, the final figure for cost of equity capital will be a judgment based on a wide variety of data and techniques. Such judgment is customary in regulatory proceedings; it is not peculiar to estimates of market-based costs of capital. The important point is that judgmental estimates of R do not have to be shots in the dark. One can arrive at rough but plausible estimates of R by using the simple tools I have just described.

5. *Econometric models*

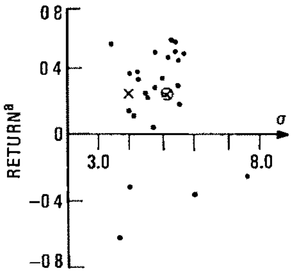
The persuasiveness of "judgmental factors" in cost of capital estimates creates a clear opportunity for effective use of econometric models. Such models may improve the accuracy of the estimates and certainly will make the required judgments more explicit.

There are many recent attempts to estimate the cost of capital via econometric techniques,[27] but the approaches taken are so diverse

FIGURE 1

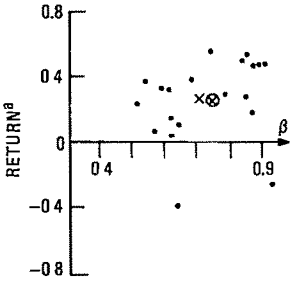
RETURN VS. RISK FOR AT&T AND STOCKS IN GROUP D

(a) UNDIVERSIFIED INVESTOR



⊗ = AVERAGE FOR GROUP D
× = AT & T

(b) WELL-DIVERSIFIED INVESTOR



[a] DEFINED AS THE AVERAGE RISK PREMIUM EARNED, 1960-1969, BASED ON MONTHLY DATA THE TREASURY BILL RATE WAS USED AS THE RISK-FREE RATE OF INTEREST FOR FULL DETAILS SEE [53], PP 49-57 AND ATTACHMENTS H THROUGH N

---

[26] *Ibid.*, pp. 49–57 and Attachments H through N. Since Figures 1(a) and 1(b) are used here only for illustrative purposes, I do not think it necessary to include all of the backup material.

[27] The list includes Miller and Modigliani [30] and the subsequent comments and elaboration on their approach [8], [17], and [42]; Gordon's work for the FCC [51], of which a new version is in preparation; also, Brigham and Gordon [7], McDonald [27], Litzenburger and Rao [26], and others.

that it is impossible to review the field here. More of a consensus will probably have to be reached before such models are routinely used. So far, Gordon's model[28] is the only one presented as evidence in a major regulatory proceeding.

5. Market efficiency and market perfection

■ There are two possible reasons for objecting to the use of the shareholders' opportunity costs in regulatory proceedings. The first is that they cannot be estimated with sufficient accuracy to support reasoned judgment by regulators. This objection cannot be answered *a priori*, but only by experience. The second objection is that data derived from stock market behavior are inappropriate for regulatory proceedings because of the market's irrationality or imperfection. This requires an answer.

☐ **Efficiency of the stock market.** Some find it difficult to rest regulatory proceedings on something so volatile as the stock market. But this volatility is basically a reflection of the fact that most assets are risky; and to say that assets are risky means that their values will fluctuate. Uncertainty is a fact of life which happens to be more dramatically disrobed in the stock market than elsewhere. Therefore, the question is not whether the stock market is stable or predictable, since part of its function is to act as a locus for risk-bearing in the economy. The question is whether the market performs this function efficiently.

An "efficient" market is one in which at any point in time security prices fully reflect all information available at that point in time, and in which prices react quickly to new information as it becomes available.[29]

Efficiency can be defined more precisely. Rewriting equation (1) with more elaborate notation,

$$P_{jt}|\Phi_t = \frac{E(\tilde{D}_{j,t+1} + \tilde{P}_{j,t+1}|\Phi_t)}{1 + E(R_j|\Phi_t)}, \qquad (1a)$$

where Φ_t is defined as a set of information. Equation (1a) defines the market's evaluation of j at t, given Φ_t. The market is efficient with *respect* to Φ_t if there is no way to use Φ_t to choose stocks with $E(\tilde{x}_{jt}) > 0$, where $\tilde{x}_{jt} = \tilde{R}_{jt} - E(\tilde{R}_{jt}|\Phi_t)$. Thus there are degrees of efficiency, depending on the breadth of information assumed included in Φ_t.

A relatively weak test is to define Φ_t as past price data and to predict that there is no superior trading strategy based on this information. Since past prices are certainly "available information," there should be no explainable price trends or cycles in an efficient market. This appears to be the case: no trading rule based on past prices has been shown to give abnormally high profits.

In fact, the evidence so far indicates that the U. S. capital markets are basically efficient with respect to a relatively broad set of information, including all data that would be regarded as publicly available. The evidence is ably summarized by Fama.[30]

---

[28] See [51].

[29] This discussion follows Fama [12].

[30] *Ibid.*

What relevance does this have for the use of finance theory in public utility regulation?

First, the market's efficiency is consistent with perfect markets—which are assumed here, and are usually built into theoretical models of valuation and the cost of capital.

Second, it indicates that observed prices at any time t approximate the equilibrium values, given Φ_t. Thus, an estimate of R at time t should be based on prices at t, not on an average of these and previous prices. There is no point in "smoothing" stock price series.

Third, market efficiency confirms that observed stock prices are closely coupled to information about the possible risks and returns of alternative investment opportunities. Otherwise a firm's "cost of capital" has little meaning or relevance. The measurement of a firm's cost of capital rests on the assumption that its stock is accurately priced relative to other equivalent-risk investments.

To summarize, there is positive evidence that overall capital market operations are basically efficient, and this efficiency is consistent with the hypothesis that market imperfections are minor.

6. Using the cost of capital as a basis for regulation

■ From this point I will assume that an estimate of Utility X's cost of capital is available. The problem is to determine how this figure should be used. I will continue to assume all-equity financing.

☐ **A straightforward approach: application to book value rate base.** I turn first to a simple and somewhat exaggerated example. Imagine a utility with book assets (rate base) of \$100 per share. It is all-equity financed. Earnings per share are \$16, all paid out as dividends. Earnings per share are expected to remain constant indefinitely.

Under such conditions, the earnings-price ratio (or the dividend yield) will measure shareholders' opportunity cost correctly. Suppose we observe a current price of \$200. Then

$$R = \frac{\text{EPS}_1}{P_0} = \frac{D_1}{P_0} = \frac{16}{200} = 0.08\,.$$

How should the regulatory commission use this information? What I will call the "straightforward approach" is to allow earnings of 8 percent on the usual book value rate base. The price of a utility's product or service will be set at a level sufficient to yield profits of 0.08 (100) or \$8.00 per share.

If investors consider the new earnings level permanent, the utility's stock price will fall to \$100:

$$P_0 = \frac{\text{EPS}_1}{R} = \frac{8.00}{0.08} = 100\,.$$

In other words, the straightforward application, in this idealized case, will drive share price to book value per share. The stock sold at \$200 in the first place only because the firm was then earning, and was expected to earn, twice the cost of capital.

A firm's market value will equal book value if it consistently earns a book rate of return equal to the cost of capital.

This example is not intended as a paradigm of ideal regulation. However, I do wish to respond to Ezra Solomon's contention that it and similar approaches are "inherently contradictory and inconsistent"[31] when there is a gap between the cost of equity capital and the book return actually being earned by the utility in question. Walter Morton has come to a similar conclusion:

> What is wrong with [the cost of capital] is that whenever it is applied to a price above book it must . . . cause a fall in earnings and a fall in the price of the stock.
>
> Any theory which postulates that investors pay above book, with the expectation that earnings will be cut as described by regulatory action, must assume either profound ignorance and ineptitude on the part of the investors, or a spirit of masochism which induces them to destroy their own capital.[32]

Actually, there is nothing inconsistent or illogical about the straightforward approach unless it is assumed that investors' expectations (as observed in the process of measuring the cost of capital) should always be confirmed when regulatory commissions act.

Take the simple example. Suppose that investors do not anticipate the rate reduction which forces earnings per share down to $8. The cost of capital is 8 percent both before and after the regulatory decision, so there can be no error or inconsistency in measurement. Rather, what happens is that the expectations investors hold before the regulatory decision (i.e., earnings per share of $16) turn out to be wrong in the event. But the regulatory commission is not bound to confirm investors' expectations. Therefore, the straightforward approach is logically sound. (Whether or not the regulators *should* force earnings per share down to $8 is, of course, another question.)

This discussion illustrates the dangers of using market value (as measured by share price) as a basis for setting earnings levels. The point is explicitly recognized in the Hope decision:

> The fixing of prices, like other applications of the police power, may reduce the value of the property which is being regulated. But the fact that the value is reduced does not mean that the regulation is invalid. . . . It does, however, indicate that "fair value" is the end product of the process of rate-making not the starting point. . . . The heart of the matter is that the rates cannot be made to depend upon "fair value" when the value of the going enterprise depends upon earnings under whatever rates may be anticipated.[33]

In short, "consistency" does not require that a market-based cost of capital must be applied to market value rate base. Actually, the problem with the straightforward approach is not one of inconsistency but of possible difficulties in measurement.

Suppose that market value is initially above book. The regulators announce that they will use the straightforward approach along with the DCF method of measurement. Share price will fall, since investors will anticipate a lower allowed rate of return after the Commission acts. On the other hand, if share price falls, the Commission will overestimate the cost of capital if they assume that investors expect continuation of past earnings. If investors, recognizing this, expect that the regulators will misread their (investors') expectations,

---

[31] Testimony of Solomon [60], transcript p. 1044.

[32] [33], p. 22.

[33] Note 1 *supra*, at 601.

TABLE 3

TEXAS EASTERN TRANSMISSION COMPANY– ACTUAL RATES OF RETURN ON EQUITY VS RATES IMPLIED BY 6.5 PERCENT OVERALL RETURN

YEAR	RATES IMPLIED BY 6.5 PERCENT OVERALL RETURN	ACTUAL RETURN	DIFFERENCE
1965	11.10	14.03	+2.93
1966	10.65	14.44	+3.79
1967	9.98	15.10	+5.12
1968	9.38	16.25	+5.87

SOURCE: FPC DOCKET RP69–13 (PHASE I), INITIAL BRIEF OF COMMISSION STAFF [59].

then clearly, we have the beginnings of a very complicated game. Moreover, there is no guarantee that the players will arrive at a solution in which R is correctly measured.[34]

The solution to this measurement difficulty is not to rely on a sample of one firm, the utility in question, but on a broader sample of equivalent-risk firms.

☐ **Will share prices be forced to book value?** In our example, straightforward application of the DCF approach forces share price down to book value per share. This is generally true if the utility can actually be expected to earn the rate set for regulatory purposes. However, this is not always a safe assumption in practice.

This can be illustrated by the actual case of Texas Eastern Transmission Company for the 1965–1968 period. The firm's rate settlement in 1965 was on the basis of an overall rate of return of less than 6.5 percent. If Texas Eastern had actually earned 6.5 percent overall from 1965 through 1968, then it would have achieved the rates of return on book equity indicated in the second column of Table 3. The actual rates of return are shown in the third column.

Although it is difficult to infer exact causes, the fact that utilities can earn more or less than is nominally allowed appears due to four factors.

1. Regulatory lag

The existence of a regulatory lag is necessary but not sufficient. That is, if prices were immediately lowered (raised) whenever a

---

[34] Suppose that there is no growth trend in the utility's sales, profits, etc., and that the regulators measure R by the earnings-price ratio. Let BV be book value per share and ROI be the book rate of return. Let the superscript ^{0} indicate initial values. Thus, in the numerical example used in this section, $\mathrm{BV} = \mathrm{BV}^0 = 100$, $\mathrm{EPS}^0 = 16$, and $\mathrm{ROI}^0 = 0.16$. The regulators are assumed to set $\mathrm{ROI} = \mathrm{EPS}^0/P$, resulting in a new earnings-per-share of $\mathrm{EPS} = \mathrm{ROI}(\mathrm{BV}^0)$. If investors recognize this, $P = \mathrm{EPS}/R$. Thus we have three equations and three unknowns, EPS, ROI, and P. Solving for ROI, we obtain

$$(\mathrm{ROI})^2 = \mathrm{EPS}^0 \frac{R}{(\mathrm{BV}^0)}$$

or

$$(\mathrm{ROI})^2 = \mathrm{ROI}^0(R).$$

Of course if $\mathrm{ROI}^0 = R$ then $\mathrm{ROI} = R$, but in this case the "game" will never begin. If the game does begin, then ROI^0 must differ from R and thus at the "solution" the new allowed ROI will not equal R either. In the numerical example, this solution will lead to $\mathrm{ROI} = 0.113$.

utility's realized return rose above (fell below) the allowed return, then the utility would always earn exactly the allowed return (assuming that there is *some* price which will generate the required profits). But the utility has an opportunity to earn more than the allowed return if regulatory surveillance is lax and/or there is delay in instituting new proceedings.

2. *Cost trends*

The tendency in regulatory proceedings is to estimate future costs per unit of output on the basis of past, or at best current, costs and output. The likely future changes in cost and output levels are not taken into account systematically. If cost trends are favorable—as a result of technological advances, for example, or of market growth when there are economies of scale—then regulatory lag will allow utilities always to stay somewhat ahead of the game.

3. *Factors not under regulatory control*

Clearly, if a utility has diversified into nonregulated fields, then restricting the profitability of the regulated portion is not sufficient to insure that the firm's book rate of return equals the cost of capital for the firm as a whole. A similar problem arises when different parts of a firm's operations are regulated by different bodies.

4. *Changes in rate base relative to capacity and output*

The size of a utility's book value rate base relative to its production capacity and output depends on the average age of its assets—that is, the older the assets the greater the proportion of the initial investment written off as depreciation. Thus, if capacity, output, and operating costs are constant, the utility's book rate of return will increase over time. The same phenomenon will occur if the utility's rate of asset expansion diminishes, other things constant.

Of course, these four factors may work against the utility as well as for it.[35]

In short, a straightforward application of the cost of capital to a book value rate base does not automatically imply that market and book values will be equal. This is an obvious but important point. If straightforward approaches did imply equality of market and book values, then there would be no need to estimate the cost of capital. It would suffice to lower (raise) allowed earnings whenever markets were above (below) book.

☐ **Mixing true and accounting rates of return.** Ezra Solomon has forcefully pointed out one major difficulty in regulatory procedures.[36] As matters stand now, regulation is based on utilities' book rates of return. The trouble is that book rates of return can be poor measures of true (DCF) rates of return. The error's direction and extent can be an extremely complicated function of the firm's growth, the average maturity of its assets, its depreciation and capitalization policy, and inflation and other factors. Solomon concludes correctly that "the rate of return in conventional book rate units is conceptually and numerically different from the rate of return in DCF units," and that "two companies with similar DCF rates of return may well show widely differing book rates of return."[37]

---

[35] What about the "differences" shown for Texas Eastern in Table 3? My understanding is that they are due to favorable cost trends and profitable diversification into nonregulated industries. One "favorable cost trend" was a reduction in the effective income tax rate due to the investment tax credit. Incidentally, the rates of return on equity implied by a 6.5-percent overall return declined because embedded debt costs rose over the 1965–1968 period.

[36] Solomon [46]. See also Solomon and Laya [47].

Clearly, this is a potentially significant problem. How should regulatory decisions respond to possible biases in book rates of return? Do biases make use of cost of capital estimates inferior to rates of return based on traditional interpretations of the comparable earnings standard?

Suppose we accept that the possible biases are very difficult to estimate and adjust for. Then the traditional interpretation has an apparent advantage of consistency, in the sense that a utility's allowed book rate of return is compared to the book returns of other firms. However, consistency would also require that the utility's performance should be compared with the performance of firms whose book rates are subject to similar biases.

This only aggravates the problems in using the traditional interpretation. The firms most likely to have similarly biased book returns are other utilities. But their returns partly reflect past regulatory actions and thus do not provide an independent standard. Book returns of unregulated firms can be used, but such firms are likely to report book returns subject to different and possibly more severe biases than utilities' returns.

The alternative is to rely on the cost of capital concept. In this case regulators are faced with the possibility that the utility's apparent (book) rate of return may be different from the true (DCF) rate of return actually being earned by the utility.[38]

Evidently difficulties exist regardless of the interpretation of comparable earnings. At the present state of the art, the possible biases just discussed above provide no grounds for preferring the traditional interpretation of the comparable earnings standard to the interpretation presented here. The matter is ripe for further research.

---

[37] [46], p. 78.

[38] Actually, the difficulties which arise when this concept is relied upon seem to be more tractable—in the long run, at least—than if the traditional approach is used. The bias need be assessed for only one firm rather than for the broad sample of firms required to implement the traditional approach. It also seems likely that the regulatory process itself restricts the bias. If a regulatory commission decides to allow a return R, and adjusts the utility's prices frequently enough that the utility always earns R on a book basis, then the utility will earn the same true return R. It must be granted that regulation does not work this perfectly. There are lags and therefore some fluctuations in book returns with unknown effects on true returns. Allowed rates of return change from time to time. Inflation is a factor. The likely effects of all these items probably cannot be assessed without building a relatively detailed simulation model. Nevertheless, I think we can anticipate the likely results of such a simulation. The biases in the book return are associated with variation in individual assets' book returns over the assets' lives. The regulatory process diminishes this variation. Since its extent is probably much less for utilities than for manufacturing companies or other unregulated firms, the bias in utilities' book returns will probably be relatively small. This argument is made in more detail by Trapnell [49].

☐ **Inflation.** A book value rate base is the original cost of assets less depreciation. Neither component is adjusted to reflect experienced inflation, which raises two questions:

(1) Does fair play require an "inflation adjustment" to either the rate base or the allowed rate of return?

(2) Is such an adjustment required for efficient allocation of resources?

I consider only the first question here and defer the second until later on in the paper.

The classic formulation of the problems of regulation in inflationary times is Walter Morton's.[39] His answer to question (1) is yes. This answer rests on a value judgment—more precisely on an operational idea of fair treatment for utility investors.[40] Unfortunately, the requirements of "fair treatment" are not clearly defined in practice.

First, note that the cost of capital, as defined here, includes an adjustment for expected inflation. Investors' opportunity costs are estimated in nominal, not real, terms. Further, since "risk" here is related to uncertainty about nominal returns, it reflects uncertainty about future inflation and its possible effects on the regulated firm. (The effects depend in turn on the responses of regulators to various degrees of experienced inflation.)

One might visualize an implicit contract between investors and regulators, specifying the regulators' response to experienced inflation among other things. The question of whether investors are being treated fairly at any point in time depends on whether the implicit contract is being honored by the regulators.

Consider again the numerical example introduced at the start of this section. The cost of capital is 8 percent and the rate base is $100. Suppose the 8-percent rate of return includes investors' expectation of 2-percent-per-year inflation.

Thus the regulators allow earnings of 0.08(100) = $8. Price is

$$P_0 = \frac{\text{EPS}_1}{R} = \frac{D_1}{R} = \frac{8}{0.08} = \$100\,.$$

Now assume that there is actually 10-percent inflation in $t = 1$, 8 percent more than expected. However, by the start of $t = 2$, investors expect inflation to drop to the "normal" 2-percent-per-year rate. Then R is again 0.08 at the start of period 2 and, according to the straightforward approach, no adjustment in the utility's allowed return is necessary. Share price will remain at $100. The utility's shareholders' wealth is $108 per share at the start of $t = 2$, including earnings paid out during $t = 1$. Obviously their real rate of return is negative.

Is this fair? It depends entirely on whether the "contract" between regulators and investors calls for investors to absorb the risks of greater-than-expected inflation. In real life, the "contract" is so vague that there is very little ground for calling any regulatory strategy fair or unfair. However, the straightforward application of the cost of capital to the book value rate base is not unfair as long as it is applied consistently.

---

[39] See [33].

[40] See [34], p. 122 *ff.*

7. Finance theory and the goals of regulation

■ In the last section I considered several possible objections to use of utilities' costs of capital as a basis for their regulation. By and large the objections can be answered. But at best this shows that regulation *can* be based on finance theory; it does not show that it *should* be. Thus we must turn to a more fundamental question: What are the goals of rate of return regulation, and to what extent can these goals be met by procedures based on finance theory?

To some extent the specification of goals is a matter complicated by the necessity for value judgements. An investor who purchases shares of Utility X at $200 in good faith will feel cheated by any regulatory decision which hands him a $100 capital loss. On the other hand, regulators cannot be bound to confirm investors' expectations in all instances.

Finance and economics are not very helpful when the problem of regulation is framed as "consumers vs. investors." Instead I will assume that regulation is intended as a substitute for competition. Thus, a "fair return" will be defined in terms of the competitive standard.

☐ **"Fair return" and the competitive standard.** Ideal regulation forces the utility to operate at competitive levels of investment, price, output, and profit.

This is difficult, perhaps impossible, to achieve in practice. Clearly, rate of return regulation can reduce or eliminate "monopoly profits;" but it is not so clear that such regulation produces the investments, outputs, or prices that would occur if a competitive solution could be achieved.

Moreover, in "naturally monopolistic" industries, with systematically decreasing average costs, a fully competitive market is not a realistic alternative. In this case, we might conclude that

> the function of regulation is to preserve for the public, instead of the producer, the benefits arising from legal monopoly without depriving the investor of a competitive profit.[41]

Thus it is natural to begin with the problem of eliminating "monopoly profits"—or, to put it more positively, the problem of providing the "fair" rate of return that would obtain in a competitive market.

What does "monopoly profits" mean? Suppose we observe an unregulated firm that has recently been very successful, one that has been able to earn more than its cost of capital. Are these high returns monopoly profits? Not necessarily: the firm may be in a competitive industry in which very high profits were not expected. (They were perhaps hoped for, but the hopes were balanced by fear of losses.) The rate of return actually being earned may be a pleasant surprise.

A superior rate of return will be a "short-run" phenomenon in competitive markets. Such a return will erode as markets shift towards long-run equilibrium. However, short-run profits or losses are more the rule than the exception. In real life the path of adjustment to long-run equilibrium will not be smooth, because of uncertainty and because the target itself will be continually changing.

In short, the theory of competitive markets provides no grounds for enforcing *ex post* equality of a utility's rate of return on assets and its cost of capital. It is more relevant to consider rates of return *ex ante.*

---

[41] *Ibid.*, p. 94.

Long-run equilibrium in competitive markets implies that the average expected rate of return on new capital investment equals the cost of capital.[42] If the average expected rate of return does not equal the cost of capital then there will be entry or exit from the industry. Thus, if the aim is to eliminate monopoly profits, this principle follows:

Regulation should assure that the average expected rate of return on desired new investment is equal to the utility's cost of capital.

This principle follows if "fair return" is understood in terms of the competitive standard. If the principle is accepted, it obviously follows that rate of return regulation should be based on finance theory and the cost of capital concept.

☐ **What "fair return" does and does not imply.** Before considering how to implement the principle, it will help to summarize some other things it does and does not imply.

(1) Note that an opportunity to invest in a project offering more than the cost of capital generates an immediate capital gain for investors. This is a windfall gain, since it is realized *ex ante.*

(2) A firm which can expect to earn its cost of capital on new investment meets the "capital attraction standard." This follows from the very definition of the cost of capital.

(3) Adherence to the principle implies that expected return to the equity owner is "commensurate with returns on investments in other enterprises having corresponding risks." Thus the principle is consistent with the comparable earnings standard established by the Hope case, provided that the standard is interpreted in terms of investors' opportunity costs.

(4) There are several things that the principle does not imply. It does not specify returns *ex post;* it is solely an *ex ante* concept. The existence of competitive markets does not require that expectations be realized for any asset, or even for all assets over any given period of time. Regulators can eliminate unexpectedly high or low rates of return after the fact, but only if they are willing to make the firm a risk-free investment.

(5) The principle says nothing about whether regulation should aim to make utilities safe or risky enterprises.

(6) Finally, it should be reemphasized that adherence to the principle does not guarantee that the utility will operate at competitive levels of price, output, and investment. This can easily be shown; it follows from the absence of any unique relationship between these variables and the *ex ante* rate of return on investment. There are many combinations of price, output, cost, and investment as well as many combinations of the various factors of production which will yield an expected rate of return equal to the cost of

---

[42] It is always true that the firm will invest up to the point where the marginal expected return on investment equals its cost of capital. This is so for both perfect competitors and monopolists, in both the short or the long run; thus, it provides no guidance for regulation.

capital. Further, there is no reason to suppose that the risk class the firm finally ends up in will be the same class that would prevail in a competitive environment.

☐ **Implementing the idea of fair return.** The principle that utilities ought to expect to earn the cost of capital on new investment is general enough to be compatible with a wide variety of regulatory schemes. Which one should be used? It is not yet possible to answer this question definitively. Nevertheless, a good deal can be said about the pros and cons of the most obvious alternatives.

In many ways the simplest approach is the "straightforward" approach with no regulatory lag. That is, the utility's product is priced at the start of each period t so that

$$REV_t = C_t + Z_t + R_t BV_t, \tag{13}$$

where

REV_t = anticipated revenues in t,
C_t = anticipated operating costs in t,
Z_t = depreciation to be charged in t,
BV_t = the rate base at the start of t—i.e., the book value of the utility's assets at that time, and
R_t = the utility's cost of capital measured at the start of t.

"No regulatory lag" means that the time from t to $t + 1$ is short enough so that deviations of actual, from anticipated, revenue and cost are not significant. Of course, it may not be easy to find a price such that equation (13) holds; both REV_t and C_t depend on output, which is in turn a function of price. But I will assume that regulators solve this problem somehow. Now consider the pros and cons of such a proposal.

Pros

First, the cost of capital will be relatively easy to measure, since a utility operating under the scheme just described will tend to be a very safe investment. The only uncertainties involve:

(1) Future changes in the cost of capital, and
(2) The possibility that there may be no price which will generate the required revenue.

It is hard to believe that an established utility facing only these uncertainties would have a cost of capital much greater than corporate bond yields. (Note that holders of corporate bonds also face the first source of uncertainty.)

Second, such a scheme would be easy to administer. The rate cases would be frequent, but routine.

Cons

There are, however, serious disadvantages. For one thing, a low cost of capital is not necessarily a good thing. There is no basis for assuming that, in a competitive market, uncertainty about operating costs would be borne almost entirely by consumers, as would be the case under this rule. Consequently, this is not likely to be an optimum allocation of risk bearing.

But the most serious item is that there is very little incentive for the utility to be efficient in choice of factor proportions, capacity, price and output, or technology. If the utility can expect to earn no more nor less than its cost of capital, then it has no incentive to seek efficiency along any of these dimensions.

It might be thought that a slight compromise of the principle—i.e., allowing the utility to expect to earn a rate R^* which is a bit greater than R—would establish the proper incentives. However, Averch and Johnson[43] have shown that the condition $R^* > R$ creates an incentive for firms to use more than the efficient amount of capital relative to other factors of production. Moreover, it is possible for the inefficiency in factor proportions to increase as the difference between R^* and R decreases.[44] Thus the compromise does not seem helpful, assuming use of the straightforward approach with no lag.

The charge of inefficiency is reinforced by Irwin Friend's argument,[45] which goes as follows. Suppose the utility is regulated by the straightforward approach, with no lag. Nevertheless, it is acting in good faith and trying to be efficient. The utility finds itself faced with a wide range of investment opportunities, some "good"—i.e., offering a rate of return greater than R—and some "bad." Efficiency would seem to call for taking only "good" projects. But this would lead to an average rate of return higher than R. Thus the utility might just as well forgo the good projects or balance them with bad ones.

In short, the straightforward approach sans lag has little to recommend it. It meets the standard of "fair return" but it accomplishes little else. In particular, it removes any incentive for efficient operating or capital budgeting procedures.

☐ **Conscious use of regulatory lag.** As I have now emphasized several times, firms in a competitive industry will not earn the cost of capital at all points in time. *Ex post* returns can deviate substantially from the cost of capital in the short run. The duration of the deviations will be limited by the time required by firms to invest (or dis-invest) and enter (or leave) the industry. A regulatory lag provides a short run in which utilities can earn unexpectedly high or low profits, and the rate proceeding at the end of the lag can play a role analogous to the forces which drive competitive industries towards long-run equilibrium.

Consider, then, regulation according to the straightforward approach but with *conscious* use of regulatory lag. I emphasize "conscious:" although there is inevitably a lag in practice, this does not necessarily imply a tolerance for surprisingly high or low rates of return. Rather, it seems to reflect a willingness to put off the next rate proceeding until profits get out of line.

With conscious use of regulatory lag, prices would be set and then left unchanged for several periods. Equation (13) would remain the starting point for determining the appropriate price. However, there are some additional complications. Suppose it turns out that trends in cost, technology, demand, etc. consistently favor the utility. Then in principle the regulators should take account of the trends. That is, revenues allowed in t would have to be lower than REV_t as given by equation (13), so that

---

[43] See [1].

[44] See Baumol and Klevorick on this point [18], pp. 175–76. The Baumol-Klevorick article reviews the extensive literature on the Averch-Johnson thesis.

[45] See Friend [50] and [37], p. 4.

$$\sum_{\tau=t+1}^{t+L} \frac{REV_\tau - \overline{REV}_\tau}{(1+R_t)^{\tau-t}} = 0\,, \tag{14}$$

where

$\overline{REV}_\tau$ = revenues expected in τ, given the price set at t;
REV_τ = revenues expected to satisfy equation (13) in τ, and
L = anticipated length of the lag.

This condition is necessary for the utility to expect to earn the cost of capital on investments undertaken in t and subsequently.

Practical difficulties

The necessity for equation (14) means that the longer the lag, the greater the administrative difficulty. To some extent there must be regulation *ex ante*, which provides for endless argument. Moreover, the utility now has the incentive to overestimate future costs. These difficulties can be ignored, but only at the expense of possibly violating the principle that utilities ought to expect to earn the cost of capital on new investment.

Another disadvantage is the difficulty in determining the appropriate duration of the lag. This cannot be left entirely to the utility, because then the lag would be short when profits are low and long when they are high. Further, it makes sense to accept unexpectedly high profits for a relatively long time if they are due to unusual managerial efficiency, but to cut the lag short if the high profits stem from the exploitation of the utility's monopoly position. At best, the straightforward approach *cum* lag could not be a formula for regulation but only an approach to it.

Effects of the regulatory lag on efficiency

The existence of a regulatory lag clearly provides the utility with an incentive to improve efficiency. The incentive appears along several dimensions:

(1) Suppose the allowed rate of return equals the cost of capital. Bailey and Coleman[46] have shown that existence of a regulatory lag will induce the utility (a) to use factors of production in efficient porportions and (b) to produce more than would an unconstrained monopoly.
(2) Existence of a lag allows the utility to capture some of the rewards of managerial efficiency and of cost-reducing innovation.[47]
(3) The lag encourages efficient capital budgeting procedures. There is a positive incentive to avoid "bad" projects offering returns less than the cost of capital, and an incentive to disengage from "bad" projects previously undertaken.

---

[46] In [2]. Bailey and Coleman also show that the Averch-Johnson effects persist when there is a lag and the allowed rate of return is *above* the cost of capital.

[47] See Baumol and Klevorick [4], pp. 182–89.

Thus, although we cannot guarantee that a straightforward approach *cum* lag will lead exactly to the competitive solution, it does move the utility in the right directions. The inclusion of a lag does not make this strategy any more or less fair, but it makes the utility more efficient.

☐ **A tentative proposal.** The implications of the discussion so far can be summarized by offering a tentative proposal. Regulators should:

(1) Determine a price for the utility's product or service such that it can expect to earn its cost of capital, given the current cost and demand functions, rate base, and scheduled depreciation.

(2) Check to see whether the utility can be expected to earn more (less) than the cost of capital in subsequent periods, given the price set at $t = 1$. If so, lower (raise) the price so that the utility can expect to earn its cost of capital over the period of the anticipated lag.

(3) Tolerate unusually high or low profits during the period of regulatory lag. The length of the lag should roughly correspond to what the short run would be if there were a competitive market—that is, the length of time necessary to adjust the amount of fixed factors of production in response to changed conditions.

Admittedly, there would be compromises in practice. Because of administrative difficulties, step (2) would probably have to be skipped except in very clear-cut cases, and any adjustment would probably be based on judgment rather than explicit use of equation (14).

This proposal differs from current procedures primarily in the conscious use of regulatory lag. It is probably not the best strategy in any ultimate sense. There are many other strategies that are consistent with the principle that utilities ought to expect to earn their cost of capital on new investment; and it will be surprising if none among these turn out to be better, at least in theory, than the straightforward approach *cum* lag. Nevertheless, this proposed approach seems attractive pending rigorous examination of alternatives.

The search for alternative regulatory strategies might proceed in any one of several directions. One open question is whether use of a book value rate base leads to the best attainable regulatory decisions. This matter is briefly reviewed in the next section.

8. Determining the appropriate rate base

■ There are basically three different concepts of rate base that could be employed. They will be abbreviated as follows:

BV	Book value, based on the usual accounting principles.
SMV	Stock market value, i.e., number of shares outstanding times price per share.
CMV	Competitive market value, i.e., whatever the utility's assets would be worth at long-run equilibrium in a competitive market.

CMV can also be defined as the original cost of the firm's assets less economic depreciation. Similarly, BV equals original cost less accounting depreciation.

Some state commissions employ a fourth concept, the "fair value" rate base; but this need not delay us. In practice, fair value seems to be defined as book value plus a modest *ad hoc* adjustment.

Thus far, all I have said is that SMV is not useful in defining a utility's rate base. There are several reasons why. First, since SMV depends on how investors expect the regulators to act, it should be the "end result. . . . not the starting point."[48] Second, adopting SMV as a rate base amounts to a commitment to confirm investors' expectations regardless of what they are based on. Third, if SMV is maintained consistently above (below) BV then the utility will expect to earn a rate of return on its new investment which is greater than (less than) the cost of capital.

But what about CMV as an alternative to BV?

☐ **Determinants of CMV relative to BV.** The concept of a CMV rate base originates in the standard theory of competitive markets. It will help to review the determinants of CMV according to this theory.

Long-run equilibrium requires that the expected rate of return on a firm's CMV be R, its cost of capital. This must be true both in an average and a marginal sense. If the marginal return on the CMV of new assets is not R, then the firm's investment decision can be improved. If the average return on the CMV of all assets is greater than R, then there will be entry of new firms; if it is less than R, capital will be withdrawn from the industry.

There are several ways to state this formally, but the most useful starts with an analogue to equation (13):

$$REV_t = \pi_t Q_t = C_t + \hat{Z}_t + R_t \mathrm{CMV}_t, \qquad (15)$$

where

$\pi_t Q_t$ = equilibrium price times quantity to be produced during the period from t to $t+1$,

$\hat{Z}_t$ = expected economic depreciation from t to $t+1$,

C_t = expected out-of-pocket cost of producing Q_t, and

R_t = cost of capital measured at t.

Long-run competitive equilibrium requires that equation (15) holds when $\pi_t = \mathrm{LRMC}_t$, where LRMC represents long-run marginal cost. For simplicity, let us omit the t's. Then the following equation may be regarded as an implicit definition of the CMV of the firm's assets:

$$\frac{R(\mathrm{CMV}) + \hat{Z} + C}{Q} = \pi = \mathrm{LRMC}. \qquad (16)$$

That is, given the equilibrium price π, each firm's CMV will adjust so that equation (16) holds. Of course competitive equilibrium also requires that π = LRAC, long-run average cost.

Under these conditions, the CMV of any new asset is simply its purchase price. Suppose a firm invests \$100 in new assets during a given year. The firm is in a competitive industry which is at long-run equilibrium. We know that the expected rate of return on the 100th dollar must be R; consequently, this marginal dollar must contribute exactly \$1 to both CMV and BV. However, the average expected rate of return on the $100 invested must also be R; otherwise there would be an incentive for capital to enter or exit from the industry. Thus the other $99 invested must contribute exactly $99 to BV and CMV. The total CMV of the firm's "old" and "new" assets is determined by equation (16).[49]

---

[48] Note 1 *supra*, at 601.

Now let us carry equation (16) over into the regulatory arena as the definition of the CMV rate base. Under what conditions will a utility's CMV and BV rate bases differ?

The answer is clearest when we consider a utility that is started "from scratch." The utility's initial rate base is its gross investment outlay. For simplicity, assume scheduled book depreciation equals expected economic depreciation $\hat{Z}$. Then

$$\text{LRAC} = \frac{C + \hat{Z} + R(\text{BV})}{Q}. \tag{17}$$

Comparing equations (16) and (17), it is clear that CMV $\gtreqless$ BV as LRMC $\gtreqless$ LRAC.

This argument reflects the essential equivalence between use of a CMV rate base and LRMC pricing. That is, one way to test whether a firm's CMV differs from its existing BV rate base is to see whether there is a difference between (1) the price derived from straightforward application of the cost of capital to BV and (2) long-run marginal cost of new capacity, based on the most efficient available technology. If price is greater than LRMC, then BV is greater than CMV, and conversely. Thus, the straightforward approach based on CMV is exactly equivalent to (long-run) marginal cost pricing.[50]

☐ **Effects of using a CMV rate base.** It is obvious that regulation by the straightforward approach, under which the utility expects to earn the cost of capital on BV, cannot be expected to lead to the competitive solution unless BV = CMV—that is, unless accumulated book depreciation approximates economic depreciation.

There are no procedural difficulties in applying the straightforward approach *cum* lag to a CMV rate base, assuming that CMV can be estimated. (The estimation of CMV is difficult, particularly for utilities, but probably not impossible.)[51] Use of CMV would call for (1) an attempt to reflect expected economic depreciation in book depreciation schedules and (2) periodic write-ups or write-downs of the rate base in response to unexpected developments.[52] However, strict use of a CMV rate base is *not* always consistent with the comparable earnings standard—that is, not always consistent with the proposition that the utility ought to expect to earn its cost of capital on new investment.

---

[49] It is *not* determined by the reproduction cost of the old assets. If "reproduction cost" is used to define rate base, it must be in the sense of providing equivalent capacity with the latest equipment and procedures. But reproduction cost so defined is simply the CMV of the old assets.

[50] Here I ignore complications introduced by regulatory lag.

[51] The special complicating factors for regulated firms include the following:

(1) If a CMV rate base is used, the utility has an incentive to overdepreciate, thereby raising its price and increasing the immediate cash return. The excess depreciation could always be made up later by an *ad hoc* write-up of assets.

(2) A CMV rate base dilutes the utility's incentive to embrace technological change. Investment in radically more efficient assets, for example, would lead to a write-down of the value of old assets.

(3) In the case of unregulated firms, SMV may be a reasonable proxy for CMV. Unfortunately, the existence of regulation breaks the link between SMV and CMV.

Consider again the case of the utility started from scratch. Suppose it is regulated according to a "straightforward approach" but with a CMV rate base. Then if LRMC is less than LRAC (the usual condition for a "natural monopoly") the utility can expect to earn less than R on its investment, since there will be an immediate write-down of the BV of this investment to CMV. Conversely, if LRMC $>$ LRAC, then the utility's shareholders will receive an immediate capital gain due to a write-up of BV. Only in the case of LRMC $=$ LRAC will this regulatory scheme adhere to the principle emphasized above, namely that a utility ought to expect to earn the cost of capital on its new investment.

It is interesting to compare this discussion with Klevorick's.[53] He found that maximization of social welfare will sometimes require that the firm be allowed to earn an average rate of return on investment that is different from the cost of capital. We have arrived at an essentially equivalent result by a somewhat more direct route. The result is that LRMC pricing (the usual condition for welfare maximization and efficient allocation of resources) will not yield an average return on investment equal to the cost of capital unless LRMC $=$ LRAC.

☐ **Conclusions.** In one sense, this is a moot result, since plausible estimates of economic depreciation or LRMC have not yet been obtained. It is not clear whether a switch to CMV rate base would require a write-up or write-down of existing BV rate bases. It may turn out that average and marginal costs are roughly equal.

The immediate implication is that more thought is required on a variety of questions. For example:

(1) Can LRMC be measured in a practical context? That is, are there administratively feasible ways to construct economic depreciation schedules or CMV rate bases?

(2) Does LRMC in fact differ from LRAC? Does application of the cost of capital to utilities' actual BV rate bases result in prices substantially different from long-run marginal costs?

(3) If the answers to question (2) are yes, might it not be possible to have your cake and eat it too (i.e., to reconcile the conflict of marginal cost pricing with the comparable earnings principle) by adopting a two-part tariff? The "use" charge would be set equal to LRMC and the "capacity" charge then adjusted so that the utility expects to earn the cost of capital on its new investment.

---

[52] The write-ups and -downs of CMV would presumably lead to corresponding changes in SMV. However, stockholders could be insulated from these changes. If they were included in ordinary income, for example, about 50 percent of the effect would be offset by extra taxes or tax shields. Alternatively, the changes could be passed on to consumers via a one-time credit or charge. It makes little difference, from an economic standpoint, which scheme is used.

[53] See [21].

9. Regulating the overall rate of return

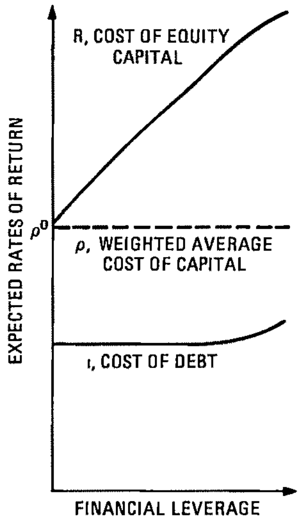
Implications of finance theory. The paper up to this point is an attempt to analyze—and to resolve, if possible—all the evident problems in applying finance theory to rate of return regulation. However, the analysis assumes that utilities are all-equity financed, which is of course not so. Thus it remains to be considered how a mixture of debt and equity financing alters the analysis. The first item is to review the theory and measurement of the cost of capital when capital structures include securities other than common stocks.

Measurement

Suppose we begin with a utility that is all-equity financed. If the utility now revises its capital structure to include some debt, the cost of equity capital (*R*) will rise, because financial leverage makes the firm's stock riskier. The interest rate, *i*, on the firm's borrowing will, of course, be less than the cost of equity capital. Figure 2 shows how *R* and *i* vary as a function of financial leverage. It should be intuitively clear that the utility's overall cost of capital, ρ, can be measured as a weighted average of *R* and *i*. It is not necessarily true that ρ will be constant, however.

FIGURE 2

EFFECTS OF FINANCIAL LEVERAGE ON COSTS OF DEBT AND EQUITY FINANCING AND THE WEIGHTED AVERAGE COST OF CAPITAL



Assume first that the firm's debt-equity mix is not expected to change. To get an exact measurement of ρ, given *R* and *i*, we can apply the same logic used to develop the cost of capital concept for the case of all-equity financing. Regardless of the degree of financial leverage, it is still possible to invest in the firm as a whole—i.e., in its assets, as distinct from any particular security. Investors can do this simply by purchasing each of the utility's financing instruments in appropriate proportions. Thus, suppose that the utility had $100 million debt outstanding (market value) and outstanding stock with a market value of $150 million. Then it would suffice to invest 40 percent of the portfolio in the utility's bonds and 60 percent in its stock.

Consider the opportunity cost of investors holding such a portfolio. The portfolio falls into a class of equivalent-risk securities and portfolios. If the expected rate of return for this class is, say, 10 percent, then the utility's stock price will adjust so that the portfolio of the utility's stocks and bonds will likewise offer an expected return of 10 percent. This 10-percent opportunity cost is the firm's overall, or "weighted average," cost of capital.

Thus, the overall cost of capital can be measured by the expected return on a portfolio of the firm's financing instruments:

$$\rho = i(D/V) + R(E/V), \tag{18}$$

where

i = current average yield to maturity of the firm's outstanding debt,
R = expected rate of return offered by the firm's stock—i.e., its cost of equity capital,
D = market value of the firm's outstanding debt,
E = aggregate market value of outstanding stock, and
$V = D + E$.

This assumes there are only two kinds of financing instruments, debt and common equity. But the weighting principle remains the same if there are others, such as preferred stock, subordinate debentures, convertible securities, etc. Of the variables used to compute the weighted-average cost of capital, only R, the cost of equity capital, is not directly observable. I have already discussed how it can be estimated.

Use of ρ as a rate of return standard

The overall cost of capital ρ is here defined as the opportunity cost of investing in a firm's *assets*. Therefore the concept and definition of ρ is the same regardless of whether the firm is financed entirely by equity or by debt and equity. The actual financing package must be taken into account in measuring ρ, but the firm's use of debt does not make measurement more difficult. Thus the conclusions reached in earlier sections of this paper are not at all dependent on the assumption of all-equity financing, provided the financing mix is taken as given. However, we face a whole new class of problems if the debt-equity ratio is expected to change. Clearly there will be a change not only in D and E, but also in R and possibly i and ρ.

Effects of changes in debt-equity ratios

The starting point in the analysis of changes in the financing mix is Modigliani and Miller's (MM's) famous Proposition I,[54] which states simply that V and ρ are constants, independent of leverage. The proposition depends on three assumptions:

(1) The existence of perfect markets;
(2) That changes in financial leverage do not affect the firm's assets, future investments, or the size and risk characteristics of its income stream; and
(3) The absence of corporate income taxes—or, alternatively, that interest charges are not tax deductable.

Figure 2 is drawn to conform with MM's Proposition I.

Let us consider the assumptions in the order stated. First, capital markets are not strictly perfect. It can thus be argued that imperfections are sufficient to negate Proposition I as an acceptable generalization. Durand's comment on MM[55] is a cogent presentation of this point of view, which is usually taken to imply that ρ is a shallow, U-shaped function of financial leverage, rather than the flat line drawn in Figure 2.

I will not discuss assumption (2) here.[56] It is not likely to be important except for firms that are levered to the point where there is a noticable probability of financial embarrassment. This is not the usual case for regulated utilities.

The tax effects are clearly substantial, however. In order to isolate them, we will assume that assumptions (1) and (2) are satisfied. Then MM's Proposition I implies, in general, that[57]

$$V = V^0 + \text{PVTS}\,, \qquad (19)$$

---

[54] Modigliani and Miller [32].

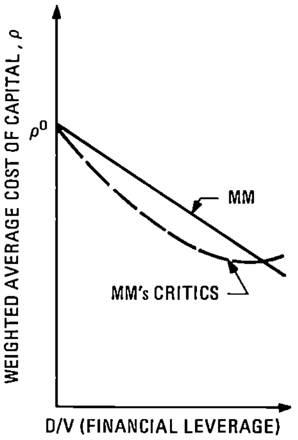
[55] See [9].

[56] The role of factors covered by this assumption is treated in Robichek and Myers [41], especially p. 16, and Baxter [5].

[57] See Robichek and Myers [41], pp. 13–15.

FIGURE 3

EFFECTS OF FINANCIAL LEVERAGE ON THE WEIGHTED AVERAGE COST OF CAPITAL WHEN CORPORATE INCOME IS TAXED



where the superscript ^{0} indicates what the value of V would be if the firm were all-equity financed, and PVTS stands for the present value of tax savings due to the deductability of interest on the firm's debt. More specifically, suppose that the firm has no growth opportunities, so that equation (6) applies. Then

$$V = \frac{X(1 + T_c)}{\rho^0} + T_c D\,, \tag{19a}$$

where

T_c = the corporate tax rate,
ρ^0 = the firm's cost of equity capital if it were all-equity financed, and
X = the firm's expected income before interest and taxes.

This implies that ρ, as defined via equation (18), is also given by[58]

$$\rho = \rho^0 - T_c(\rho^0 - i)D/V\,. \tag{20}$$

The implied behavior of ρ vs. D/V is shown in Figure 3 by the downward-sloping solid line. To repeat, this assumes that MM's proposition I holds exactly except for tax effects—i.e., that the line would be horizontal in the absence of such effects.

If MM's critics are right, the true relationship is like the dashed line in Figure 3. Unfortunately, there is no theory for this case specifying an exact functional relationship comparable to equation (20). Therefore I will continue to use MM's equations for purposes of discussion.

Equation (20) applies regardless of whether the firm is regulated or not. This may seem surprising in view of the treatment of taxes in regulatory proceedings. The usual procedure is to treat a utility's tax bill as an operating cost. If taxes change, the price of the utility's product or service is adjusted to provide an offsetting change in revenue. Thus, if a utility issues more debt, thereby reducing taxes relative to equity financing, the tax savings will be passed on to consumers in the form of lower prices. However, this does not change either equation (19) or (20); but it does make X, earnings before interest and taxes, a function of D. Specifically, for a regulated firm,[59]

$$X = X^0 - T_c iD\,, \tag{21}$$

where X^0 equals earnings before interest and taxes if the utility is all-equity financed. For an unregulated firm, $X = X^0$.

To put it another way, both regulated and nonregulated firms benefit from debt via a lower cost of capital, ρ. But unregulated firms benefit further in that X is *not* decreased to offset tax savings, as is (presumably) the case for regulated firms.

☐ **Practical implications of possible changes in financial leverage.** The implications of all this can be summed up in Figure 3, in which the solid line represents equation (20) and the dashed line incorporates the effects of the market imperfections which MM's critics think are important.[60] However, there are implications for regulation regardless of which view is correct. One point is that a measurement of ρ on the basis of an observed capital structure will be in error if investors expect capital structure to change.

---

[58] See Modigliani and Miller [31], p. 439.

[59] See Elton and Gruber [10] for a more detailed discussion of taxes and the regulated firm's cost of capital.

What is the possible magnitude of the error? Suppose equation (20) is right, and that a utility announces that it will shift D/V by 10 percent. Suppose that $T_c = 0.5$ and $\rho^0 - i = 0.05$. Then ρ will change by $T_c(\rho^0 - 1)\Delta D/V = 0.5(0.05)0.1 = 0.0025$, or 0.25 percent. In a practical context this is not a very large number, for several reasons:

(1) If MM's critics are right, and the utility's D/V ratio puts it in the trough of the dashed U-shaped curve in Figure 3, then ρ will change by less than equation (20) would indicate.

(2) It could easily take several years for a firm to shift its capital structure by 10 percent.[61]

(3) Any error in measuring ρ will persist only during the regulatory lag.

(4) To be sure, ±0.25 percent may amount to a lot of dollars. But it would explain only a fraction of the differences in the proposed costs of capital presented by the various parties in an actual regulatory proceeding.

To summarize, it does not seem too dangerous to estimate a utility's cost of capital on the assumption of a constant debt-equity ratio. Although a planned, major, rapid shift in capital structure should be taken into account, such shifts are presumably more the exception than the rule.

☐ **Regulation of capital structure.** Now we may turn to a different problem: Should utilities' capital structures be regulated? At first glance, the objective of regulating capital structure would seem to be minimization of the weighted-average cost of capital. However, it is not at all clear that this is a good thing.

Suppose MM are right, so that equation (20) holds. Then ρ will decline with financial leverage, but only because the effect of the tax subsidy to debt financing is to reduce the risk of a portfolio of the firm's stocks and bonds. However, the firm's assets are no more nor less risky: It is simply that a part of the risk is absorbed by "we the people" via fluctuations in the corporate income tax revenues. Why should this be an objective of national economic policy?

If MM's critics are right, then ρ may also decline because of market imperfections. That is, the market imperfections will lead investors to prefer corporate debt to debt undertaken on personal account, and investors will therefore require, *ceteris paribus*, a lower expected rate of return on investments in levered firms. If this is the case, it is desirable for utilities to provide corporate leverage, since investors will thereby consider themselves better off. The trouble is that we do not know how strong investors' preference for corporate vs. personal leverage really is, if indeed this preference exists at all.

---

[60] The following is a sampling of recent empirical studies, which contain references to earlier studies: Miller and Modigliani [30]; the "comments" on this piece [8], [17], and [42]; and Brigham and Gordon [7]; but see also Elton and Gruber [10], Sarma and Rao [43], and Litzenberger and Rao [26].

[61] For example, AT&T relied almost exclusively on debt for new external financing during the period 1967–1971. The effect was to shift its ratio of debt to total (book) capitalization from 33 to 45 percent. This was during a period of very heavy requirements for external finance. See Scanlon [54].

Thus, on purely economic grounds, the argument for regulating capital structure seems weak at best. About all that can be said is that utilities ought not to borrow so much that their solvency is endangered.

☐ **Some additional considerations.** The alert reader will have noticed two important differences between the overall cost of capital, as given by equation (18), and the procedure actually used by regulatory bodies in arriving at an overall rate of return allowance.[62] The differences are that

(1) Market value weights are used in equation (18), whereas book value weights are used in practice, and

(2) Embedded debt costs are used in practice.

What does finance theory say about the effects of these practices?

Clearly, the fact that the cost of capital can be applied to a book value rate base does not mean that book weights should be used in measuring it. The definition of the cost of capital in terms of investors' opportunity costs definitely implies that market value weights should be used.

This is not the whole story, however. Suppose Utility X is partly debt financed. Its regulators estimate ρ as 10 percent and therefore set X's prices so that the firm can expect to earn 10 percent on its book value rate base. X's stock sells for $100 per share after all this is done. Now suppose that interest rates rise by 2 percent, and the firm's overall cost of capital rises from 10 to 12 percent. This leads to another rate hearing in which the utility is allowed 12 percent rather than 10.

Given the regulators' response, the increase in interest rates will not affect the total market value of the firm, $V = D + E$. However, bondholders will suffer a capital loss, which implies that shareholders will receive a capital gain. The rise in interest rates will lead to a capital gain on the initial share price of $100. Conversely, if interest rates fall, bondholders will gain at the expense of stockholders.

Now, suppose that regulators wish to prevent stockholders from gaining when interest rates rise, and also wish to protect them from loss when interest rates fall. The way to do this is by allowing the firm to earn

(1) Its actual embedded interest cost, plus

(2) Equity earnings equal to the cost of equity capital times the book value of equity.

Under this procedure the overall rate of return will be a weighted average of embedded debt costs and the cost of equity capital, using book weights.

---

[62] See p. 61 *infra.*

To summarize, a regulatory strategy based on current debt costs requires market weights to compute the overall cost of capital. A strategy based on embedded debt costs requires use of book weights to determine the desired rate of return allowance. In general, the desired rate of return allowance in the latter case will not equal the overall cost of capital; that is, it will not generally reflect the current opportunity cost of investing in the firm's assets.

It is probably simpler, conceptually, to forget about embedded debt costs and simply use current borrowing costs. However, it is not necessarily illogical to use embedded costs. In fact, if there is no regulatory lag the use of embedded costs is generally consistent with the principle that utilities ought to expect to earn the cost of capital on new investment.[63] If utilities could change their rates automatically any time their embedded debt cost changed, then the rate of return earned on new investment would always reflect the actual current cost of any debt financing associated with the new investment.

But suppose there is a lag. Then if interest costs rise above embedded costs, a utility's rate of return on new investment will tend to be less than the cost of capital. The converse occurs when interest rates fall.

Conclusions

The use of embedded debt costs and book weights is not the most logical procedure. It seems simpler to rely on market weights and current interest costs, particularly when there is conscious use of regulatory lag. However, the straightforward approach *cum* lag is not necessarily incompatible with embedded debt costs and book weights, since the approach can be restricted to the equity component of utility capitalization. However, if the difference between current and embedded costs leads to a substantial violation of the principle that utilities ought to expect to earn the cost of capital in new investment, then an *ex ante* adjustment should be made.

10. Conclusions

■ This paper was motivated by dissatisfaction with the traditional interpretation of the comparable earnings standard, and by the hope that regulation could be made more effective by greater reliance on finance theory. Specifically, finance theory suggests that the comparable earnings standard should be defined in terms of investors' opportunity costs—i.e., in terms of utilities' costs of capital. Thus the paper is addressed to the question of whether regulation can or should be based on finance theory and the cost of capital concept. As we have seen, the search for an answer to this question requires consideration of a wide variety of topics in finance and the law and economics of regulation.

Is it in fact reasonable to base rate of return regulation on finance theory? It seems to me that the answer is a tentative yes. It is tentative because there are difficulties (though the difficulties are often shared by other approaches to regulation) and because it is always possible that this paper will be made moot or obsolete by future developments in law, economics, or the conditions facing regulated firms.

---

[63] The timing of the change to current interest rates and market weights is a delicate matter. As this is written, current interest costs are substantially (1.5 percent or more) above embedded costs. A switch at this time would lead to a substantial capital gain for utility investors, financed by a one-time price increase. This raises problems of equity (to say nothing of politics), because it results from a change in the rules and is not a consequence of the regulatory process as consumers or investors have understood it.

Nevertheless there is a strong positive case to be made. I will close by summarizing it briefly:

(1) Regulation is usually regarded as a substitute for competition. If the definition of "fair return" is based on the theory of competitive markets, then regulation should assure that the average expected rate of return on new utility investment is equal to the utility's cost of capital. This is the ultimate justification for basing rate of return regulation on finance theory.

(2) This principle is consistent with the comparable earnings standard—in fact, more directly consistent than the traditional interpretation.

(3) There are many ways in which the principle can be implemented. Only the most obvious, "straightforward" approach was investigated here. But this approach is logically sound and practical. By and large, the objections to straightforward approaches can be answered satisfactorily.

(4) The straightforward approach is amenable to conscious use of regulatory lag. This does not make the approach more or less fair, but it does create stronger incentives for efficiency.

(5) The cost of capital is not directly observable, since it is defined in terms of investors' expected rates of return. At the present state of the art any estimate is part judgment. However, plausible estimates can be obtained, and the need for judgment is not evidently greater when regulation is based on finance theory rather than on traditional procedures.

The straightforward approach *cum* lag is not the final answer. It is only one of many ways to implement the principle that a utility ought to be able to expect to earn its cost of capital on new investment. Probably some of the other ways are better; it is hard to believe that the usual book value rate base could not be improved upon, for example. As a matter of fact, the whole existing framework of rate of return regulation, which was taken as given for purposes of this paper, may not be best. But all of this awaits further work.

References

1. Averch, H. and Johnson, L. L. "Behavior of the Firm Under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1969), pp. 1062–69.
2. Bailey, E. E. and Coleman, R. D. "The Effect of Lagged Regulation in the Averch-Johnson Model." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 278–92.
3. Baumol, W. J. "Reasonable Rules for Rate Regulation: Plausible Policies for an Imperfect World," in *The Crisis of the Regulatory Commissions*, P. W. MacAvoy, ed., New York: W. W. Norton and Co., 1970.
4. ———, and Klevorick, A. K. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–90.

5. Baxter, N. D. "Leverage, Risk of Ruin and the Cost of Capital." *Journal of Finance*, Vol. 22, No. 4 (September 1967), pp. 385–404.
6. Black, F. and Scholes, M. "Dividend Yields and Common Stock Returns: A New Methodology." Unpublished Working Paper, Sloan School of Management, Massachusetts Institute of Technology, 1971.
7. Brigham, E. and Gordon, M. J. "Leverage, Dividend Policy and the Cost of Capital." *Journal of Finance*, Vol. 23, No. 1 (March 1968), pp. 85–104.
8. Crockett, J. and Friend, I. "Some Estimates of the Cost of Capital to the Electric Utility Industry 1954–57: Comment." *The American Economic Review*, Vol. 59, No. 5 (December 1967), pp. 1258–67.
9. Durand, D. "The Cost of Capital, Corporation Finance, and the Theory of Investment: Comment." *The American Economic Review*, Vol. 49, No. 4 (September 1959), pp. 639–55.
10. Elton, E. J. and Gruber, M. J. "Valuation and the Cost of Capital in Regulated Industries." *Journal of Finance*, Vol. 26, No. 3 (June 1971), pp. 661–70.
11. Evans, J. L. and Archer, S. H. "Diversification and the Reduction of Dispersion: An Empirical Analysis." *Journal of Finance*, Vol. 23, No. 5 (December 1968), pp. 761–67.
12. Fama, E. F. "Efficient Capital Markets: A Review of Theory and Empirical Work." *Journal of Finance*, Vol. 25, No. 2 (May 1970), pp. 383–417.
13. Fisher, L. and Lorie, J. H. "Rates of Return on Investments in Common Stock: The Year-by-Year Record, 1926–65." *Journal of Business*, Vol. 41, No. 3 (July 1968), pp. 291–316.
14. Friend, I. and Blume, M. "Measurement of Portfolio Performance Under Uncertainty." *The American Economic Review*, Vol. 60. No. 4 (September 1970), pp. 561–75.
15. Friend, I. and Puckett, M. "Dividends and Stock Prices." *The American Economic Review*, Vol. 54, No. 4 (September 1964), pp. 656–82.
16. Gordon, M. J. *The Investment, Financing and Valuation of the Corporation.* Homewood, Illinois: Richard D. Irwin, Inc., 1962.
17. ———. "Some Estimates of the Cost of Capital to the Electric Utility Industry 1954–57: Comment." *The American Economic Review*, Vol. 57, No. 5 (December 1967), pp. 1267–78.
18. Hamada, R. J. "The Effects of Leverage and Corporate Taxes on the Shareholders of Public Utilities," in *Rate of Return Under Regulation: New Directions and Perspectives*, H. M. Trebing and R. H. Howard, eds., East Lansing Michigan: Institute of Public Utilities, Michigan State University, 1969.
19. ———. "Portfolio Analysis, Market Equilibrium and Corporate Finance." *Journal of Finance*, Vol. 24, No. 1 (March 1969), pp. 13–32.
20. Jensen, M., ed. *Studies in the Theory of Capital Markets.* New York: Praeger Publishers, 1972.
21. Klevorick, A. K. "The 'Optimal' Fair Rate of Return." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 122–53.
22. Leventhal, H. "Vitality of the Comparable Earnings Standard for Regulation of Utilities in a Growth Economy." *Yale Law Journal*, Vol. 74, No. 6 (May 1965), pp. 989–1010.
23. Lintner, J. "Expectations, Mergers and Equilibrium in Perfectly Competitive Securities Markets." *The American Economic Review*, Vol. 61, No. 2 (May 1971), pp. 101–11 .
24. ———. "Security Prices, Risk and Maximal Gains from Diversification." *Journal of Finance*, Vol. 20, No. 5 (December 1965), pp. 587–613.
25. ———. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *The Review of Economics and Statistics*, Vol. 47, No. 1 (February 1965), pp. 13–37.
26. Litzenberger, R. L. and Rao, C. U. "Estimates of the Marginal Rates of Time Preference and Average Risk Aversion of Investors in Electric Utility Shares: 1960–66." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), p. 265–77.
27. McDonald, J. G. "Required Return on Public Utility Equities: A National and Regional Analysis: 1958–1969." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Autumn 1971), pp. 503–14.
28. Markowitz, H. "Portfolio Selection." *Journal of Finance*, Vol. 7, No. 1 (March 1952), pp. 77–91.

29. ———. *Portfolio Selection: Efficient Diversification of Investments.* New York: John Wiley and Sons, Inc., 1959.
30. Miller, M. and Modigliani, F. "Some Estimates of the Cost of Capital to the Electric Utility Industry: 1954–57." *The American Economic Review*, Vol. 56, No. 3 (June 1966), pp. 333–91.
31. Modigliani, F. and Miller, M. H. "Corporate Income Taxes and the Cost of Capital: A Correction." *The American Economic Review*, Vol. 53, No. 3 (June 1963), pp. 433–43.
32. ———. "The Cost of Capital, Corporation Finance and the Theory of Investment." *The American Economic Review*, Vol. 48, No. 3 (June 1958), pp. 261–97.
33. Morton, W. A. "Guides to a Fair Rate of Return." *Public Utilities Fortnightly*, Vol. 86, No. 2 (July 2, 1970), pp. 17–30.
34. ———. "Rate of Return and the Value of Money in Public Utilities." *Land Economics*, Vol. 28 , No. 2 (May 1952), pp. 91–131.
35. Mossin, J. "Equilibrium in a Capital Asset Market." *Econometrica*, Vol. 34, No. 4 (October 1966), pp. 768–83.
36. Myers, S. C. "Procedures for Capital Budgeting Under Uncertainty." *Industrial Management Review*, Vol. 9, No. 2 (Spring 1968), pp. 1–20.
37. New York Public Service Commission. *New York Telephone Co.* Prepared Testimony, I. Friend. 1971.
38. Posner, R. A., "Taxation by Regulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 22–50.
39. Rao, C. U. and Litzenberger, R. H. "Leverage and the Cost of Capital in a Less Developed Capital Market: Comment." *Journal of Finance*, Vol. 26, No. 3 (June 1971), pp. 77–82.
40. Robichek, A. A. and Myers, S. C. "Conceptual Problems in the Use of Risk-Adjusted Discount Rates." *Journal of Finance*, Vol. 21, No. 5 (December 1966), pp. 727–30.
41. ———. "Problems in the Theory of Optimal Capital Structure." *Journal of Financial and Quantitative Analysis*, Vol. 1, No. 2 (June 1966), pp. 1–35.
42. Robichek, A. A., McDonald, J. G., and Higgins, R. C. "Some Estimates of the Cost of Capital to the Electric Utility Industry 1954–57: Comment." *The American Economic Review*, Vol. 57, No. 5 (December 1967), pp. 1278–88.
43. Sarna, L. V. L. N. and Rao, K. S. H. "Leverage and the Value of the Firm." *Journal of Finance*, Vol. 24, No. 4 (September 1969), pp. 673–77.
44. Sharpe, W. F. "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk." *Journal of Finance*, Vol. 19, No. 4 (September 1964), pp. 425–42.
45. ———. *Portfolio Theory and Capital Markets.* New York: McGraw-Hill Book Co., 1970.
46. Solomon, E. "Alternative Rate of Return Concepts and Their Implications for Utility Regulation." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 65–81.
47. ———, and Laya, J. E. "Measurement of Company Profitability: Some Systematic Errors in the Accounting Rate of Return," in *Financial Research and Management Decisions*, A. A. Robichek, ed., New York: John Wiley and Sons, Inc., 1967.
48. Stigler, G. J. "The Theory of Economic Regulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 3–21.
49. Trapnell, G. R. "Relationship of the Book Rate of Return on Investment to the Internal Rate of Return on the Investment and Implications for Utility Regulation." Unpublished Manuscript. Staff Paper 4, Office of the Actuary, Social Security Administration, 1971.
50. U. S. Federal Communications Commission. *American Telephone and Telegraph Co.* Prepared Testimony, I. Friend. F.C.C. Docket 16258, 1966.
51. ———. *American Telephone and Telegraph Co.* Prepared Testimony, M. J. Gordon. F.C.C. Docket 16258, 1966.
52. ———. *American Telephone and Telegraph Co.* Prepared Testimony, D. Kosh. F.C.C. Docket 19129, 1971.
53. ———. *American Telephone and Telegraph Co.* Prepared Testimony, S. C. Myers. F.C.C. Docket 19129, 1971.
54. ———. *American Telephone and Telegraph Co.* Prepared Testimony, J. J. Scanlon. F.C.C. Docket 19129, 1971.

55. ———. *Communications Satellite Corp.* Prepared Testimony, E. F. Brigham. F.C.C. Docket 16070, 1971.
56. U. S. FEDERAL POWER COMMISSION. *Area Rate Proceeding* (Offshore Southern Louisiana . . .) *et al.* Prepared Testimony, E. Solomon. F.P.C. Docket AR 69-1 et al., 1969.
57. ———. *Consolidated Gas Supply Corp.* Prepared Testimony, H. G. Roseman. F.P.C. Docket RP69-19, 1969.
58. ———. "Opinion and Order Determining Fair Rate of Return." F.P.C. Opinion No. 579, *Pacific Gas Transmission Corp,* F.P.C. Docket RP70-4 1970.
59. ———. *Texas Eastern Transmission Corp.* Initial Brief of Commission Staff. F.P.C. Docket RP69-13, 1969.
60. ———. *Texas Eastern Transmission Corp.* Prepared Rebuttal Testimony, E. Solomon. F.P.C. Docket RP69-13, 1969.
61. ———. *Texas Eastern Transmission Corp.* Prepared Testimony, S. C. Myers. F.P.C. Docket RP69-13, 1969.
62. WILLIAMS, J. B. *The Theory of Investment Value.* Amsterdam: North-Holland Publishing Co., 1968.

The appropriate scope of regulation in the cable television industry

Richard A. Posner

Professor of Law
University of Chicago

The development of cable television has been subjected to a comprehensive and complex array of federal, state, and local regulations; and many new proposals for regulation are being discussed. This paper attempts a critique of the emerging pattern of regulation. The author analyzes the major policy choices and concludes that only limited regulation can be justified. He proposes a new federal statute that would carefully delimit the respective roles of federal, state, and municipal regulators of cable television.

1. Introduction

■ The government has not left development of cable television to the free market. Extensive regulation of rates, programs, ownership, and other facets of cable television service has been proposed and, in large measure, already implemented. A critique of the emerging pattern of regulation would seem to be timely.

I limit discussion here to the fundamental policy choices. To expound the details of particular regulations and proposals, most of which are in the nature of ephemeral and unprincipled comprises, would serve only to obscure the basic issues. Section 2 of this paper deals with a theme of growing importance in a number of substantive areas of cable television regulation: the use of regulation to compel the provision of services that do not pay their way in the market—what I call "taxation by regulation." Sections 3–5 analyze the substantive areas of regulation, divided into three broad groups: (1) regulation concerned with the content of programming on cable television; (2) regulation concerned with the level of the rates charged to the subscribers by cable television companies; and (3) regulation designed to reduce the impact on the television broadcasting industry of the competition of cable television. Section 6 considers the in-

---

Richard A. Posner received the A.B. from Yale University and the LL.B. from Harvard Law School. He is a Research Associate of the National Bureau of Economic Research and editor of *The Journal of Legal Studies*, a new journal published by the University of Chicago Law School. Professor Posner's interests include the application of economic theory to law, especially in the regulated industries, and the empirical study of the legal system.

This paper was commissioned by the Sloan Commission on Cable Communications. Section 4 is an abridgment of Posner, "Cable Television: The Problem of Local Monopoly" [23], and Sections 3 and 5 draw in part on research supported by The Rand Corporation under a grant from the Ford Foundation. In addition, ideas expressed in this paper were developed in part in the Workshop in Regulation of Economic Activity of the University of Chicago Law School, supported by The Brookings Institution. None of the views or conclusions expressed in this paper should be interpreted as reflecting the policies or opinions of the Sloan Commission, the Ford Foundation, The Rand Corporation, or The Brookings Institution.

stitutional problems created by the fact that three levels of government—municipal, state, and federal—are vying to regulate the industry, and summarizes the article's major recommendations.

2. Taxation by regulation

■ Regulatory agencies often use their control over the rate structure of regulated firms to compel the provision of unremunerative services, in much the same way that the taxing and spending powers of legislatures are used to compel the provision of services that the market would not provide, at least in the desired quantity.[1] The basic mechanism is the "internal subsidy," whereby a firm (or group of firms) provides some services at a loss which it recoups by charging a price above cost for other services. Internal subsidization on a systematic basis is possible only in a regulated industry or in one, like the postal service, where the government is the provider of the service. In an unregulated private market, firms would not provide unremunerative services for any length of time.

Internal subsidies have been quite common in the communications field[2] and already have a considerable foothold in the cable television industry. Municipalities, as a condition of franchising cable television companies, commonly require that the franchisee provide one or more channels, free of charge, to the licensing authority for public purposes. The FCC insists that the larger cable companies, as a condition of being permitted to carry distant signals, provide some original programming[3]—a stipulation that would be superfluous if such programming were expected to be profitable—and is toying with the idea of requiring cable operators to remit 5 percent of their subscription revenues to noncommercial broadcasters. Additional instances of internal subsidization in the cable industry include New York City's public-access channels[4] and occasional suggestions for wiring up slum urban areas at preferential rates.

A recent paper by John McGowan, Roger Noll, and Merton Peck provides a clue to the potential importance of the practice.[5] Their procedure is to quantify the "surplus" (approximately the potential monopoly profits) from the sale of those cable communications services for which there is a market and to appraise the claims of various interests to the surplus. Such an approach assumes that internal subsidization is the crucial issue in cable television regulation.

Despite wide acceptance of the practice, I am persuaded that it would be a mistake to countenance it in the cable industry, not least from the standpoint of those who wish to see cable service rapidly attain a very high degree of market penetration. In order to obtain a fund out of which to defray the expense of unprofitable services, the industry would have to fix rates substantially above cost to subscribers (their paying customers). The McGowan, Noll, and Peck study assumes that suppliers of cable television service can exact monopoly prices from subscribers. It is the resulting monopoly profits that would be the source of the fund. Yet the study does not regard monopoly pricing as an inevitable condition of the industry. On the contrary, it assumes that regulation could be employed effectively to compress price to cost, thereby securing the entire surplus for consumers. One may have less faith in the efficacy of regulation than does the study, yet agree (for reasons discussed later) that monopoly pricing of cable service is not inevitable.

---

[1] See Posner [26].

[2] Some examples: Telephone rates display a uniformity at sharp variance with the structure of costs, and telephone company officials have often defended the rate structure by reference to the alleged social desirability of internal subsidization. The reader may recall that when it first appeared that communications satellites might be able to distribute network television programs at highly remunerative rates, the Ford Foundation proposed that a portion of the profits be reserved for noncommercial television—by definition a service not provided in the market at a price that covers costs. The proposal is discussed in Dirlam and Kahn [8].

[3] See First Report and Order in Docket 18397 (CATV) [32]. The requirement was recently invalidated by a federal court of appeals in a case that the Supreme Court has agreed to review: Midwest Video Corp. v. United States, 441 F. 2d 1322 (8th Cir. 1971).

[4] See New York Mayor's Advisory Committee Report [18].

[5] See [16].

By either route, it follows that a policy of internal subsidization is likely to lead to higher subscriber rates than alternative policies.[6] This in turn will result in a reduction in the scale of the industry, for the demand for cable television cannot be completely inelastic, especially given the over-the-air alternatives. There may also be serious dynamic consequences. The McGowan, Noll, and Peck study apparently assumes that whether the surplus that remains after all internal subsidies have been paid is one dollar or one billion dollars, the expansion of the cable industry will be unaffected; but this is dubious. If the entire surplus accrues to subscribers in the form of lower prices for cable service, more households will subscribe sooner. Notice that internal subsidization does not infuse an industry with new resources for growth. It retards development in some submarkets (by compelling higher prices in them) in order to accelerate it in others. Where the burdened service is basic to the industry, as is getting the wire into the home, internal subsidization seems especially unlikely to improve the rate and direction of the industry's growth, and may retard it.[7]

Internal subsidization also erects a barrier to the displacement of old technology by new. One may illustrate with reference to the Ford Foundation proposal for the establishment of a nonprofit corporation to distribute network television programming by satellite,[8] with the profits of the corporation going to support noncommercial television. Suppose that the proposal has been adopted and several years later AT&T perfected a terrestrial laser system for distributing programs at lower cost. The networks would want to switch to AT&T. Noncommercial television, having meanwhile grown dependent on the revenues from the satellite system, could be expected to press for a rule either forbidding AT&T to offer the substitute service or compelling it to price the service at a level that would preserve a large share of the market for the satellite system.

---

[6] If monopoly pricing were inevitable, one effect of internal subsidization would be a transfer of monopoly profits from the stockholders of cable companies to the recipients of the subsidized services. But it would not be the only effect. If the subsidy were collected by a tax on costs or prices, the firm would raise price and reduce output. See Posner [23]. Proposals for internal subsidies of cable service, such as a 5-percent gross-receipts tax on cable companies for public broadcasting, are of this kind.

[7] See Coase [3].

[8] See note 2 *supra*.

The cable television industry presents an analogy. In view of the rapidity with which communications technology has been changing, one cannot assume that the cable represents the last word in the distribution of broadband communications. Satellites broadcasting directly to household television receivers may someday provide the same service as cable television systems at lower cost. Such a development would be retarded if an assumed surplus from cable service had been earmarked for specified public services, whose providers and beneficiaries would bring pressure to bear on government to prevent the erosion of the surplus by a new competitor.

At the very least, the possibility of new competition necessitates a cumbersome administrative apparatus to control entry. The last example assumed that direct satellite broadcasting might someday be a cheaper alternative to cable television. Suppose it is not cheaper; it might seem cheaper to the consumer—the price of the service might be less than the cable subscription rate—not because the cost was lower but because the subscription rate had built into it a charge for defraying a subsidy elsewhere in the system. To prevent diversion of business to the apparently cheaper but really dearer service requires public control over entry.

Let us now shift attention from efficiency to equity. Internal subsidization is a method of redistributing income. The subscriber who pays a higher rate than if the entire "surplus" were passed on to the subscribers is put in a worse position; whoever benefits from the subsidized services is put in a better position. One of the characteristics of internal subsidization as a method of redistribution is that it is likely to be regressive. It is essentially an excise tax—on a service, in this instance, that cannot be regarded as a "luxury" item. Enabling as it does the substitution of "free"[9] service for non-free sources of information and entertainment such as newspapers and motion pictures, television—cable television where over-the-air signal reception is very poor or the number of available over-the-air signals are very few—is well nigh indispensable to lower-income families.

If cable revenues are used to support public broadcasting, the result will be a redistribution of wealth from a lower to a higher income group, because noncommercial programming (with a few exceptions like "Sesame Street") is geared to highly educated, well-to-do viewers. If cable television revenues were used to finance the extension of cable service at zero or very low rates to slum neighborhoods, then the redistribution might be in the opposite direction. But even here one must be careful. The additional cost involved in complying with the requirement might retard the spread of cable service and this could delay service to all, including many lower-income people who would have been willing to pay marginal cost charges to obtain cable service.

There is a final point: The methods used in the McGowan, Noll, and Peck study to project the "surplus" of revenue over cost in the cable television industry permit only the crudest approximation. One cannot make accurate forecasts of aggregate nationwide consumer demand for a service that barely exists in major television markets. The resulting uncertainty, however, must greatly complicate the administration of any program of internal subsidization. Consider the FCC's proposed 5-percent cable television "dividend" for public broadcasting. If the surplus is less than at present imagined, such an exaction could have a debilitating effect on the industry's development;[10] if it is larger it could produce a revenue for public broadcasting that was in excess of the legitimate needs of that service when balanced against other claims on available sources of funds. Uncertainty as to the likely support forthcoming would retard planning by the recipients of the subsidy.

---

[9] Over-the-air television is not strictly free: its cost is presumably reflected in the prices for the products advertised on television. But this method of financing the cost of broadcasting is highly favorable to the poor in comparison with charging viewers directly. See Lees and Yang [14]. Furthermore, as discussed *infra*, p. 105, it is not even clear that the prices of products now advertised on television would be lower were television advertising forbidden.

3. Regulation of cable television programming

■ Much of the sentiment for internal subsidization of cable service reflects dissatisfaction with the program policies of broadcast television and a fear that without regulatory action the program fare carried on cable television will be no more satisfactory. We consider first whether this fear has a basis.

☐ **The impact of cable television on the quality or diversity of programming.** A well-functioning market may be assumed to provide the type and amount of a good that best satisfies the tastes of consumers, aggregated according to willingness to pay. In the case of over-the-air television service, however, the scarcity of channel space may appear to destroy any presumption that the market would satisfy a wide variety of tastes—and the abundance of channel space on the cable, to restore it. The matter, I believe, is more complicated.

There is a good deal of unused over-the-air television channel space in the UHF band today. Many channels allocated to UHF broadcasting by the FCC have never been assigned: no one has stepped forward to apply for a broadcasting license. Chicago, Detroit, and Philadelphia are among the many major markets where there is an unassigned commercial UHF channel. Furthermore, many independent UHF stations do not carry anywhere near a full week's programming, but come on the air in the late afternoon and go off before midnight.

A likely reason for so much unused air time is that the market for air time is thin. There are, to be sure, alternative explanations. One is that the UHF signal is weak, limiting audience and, hence advertiser, demand. In fact, UHF broadcasters are authorized to transmit as strong a signal as the VHF broadcasters (UHF channel 20 in Washington, D. C., for example, has the most powerful television transmitter in the entire country). Many UHF signals are indeed understrength, but only because the additional cost of maintaining a stronger signal is deemed to exceed the additional advertising revenues (resulting from a larger audience) that such a signal would generate. Another possibility is that the cost of over-the-air broadcasting is very high. According to the McGowan, Noll, and Peck study, however, the cost is approximately the same as that of broadcasting on a channel of cable television.[11] The cost of transmission is, in any event, dwarfed by the talent and other costs incurred in producing a television program.

---

[10] A recent and responsible study concluded that the cost to the cable industry of complying with the various requirements imposed or proposed to be imposed by the FCC was much higher than anyone—or at least the FCC—had believed. See Comanor and Mitchell [6].

Two explanations for the large amount of unused air time that cannot be so readily rejected are, first, the fact that even today only one out of two television sets have UHF reception, and, second, the absence of click tuning for UHF, which, silly as it may seem, is widely thought to deter viewers from watching UHF stations. These are transitory problems. All new television receivers are required by law to provide UHF reception and the FCC will soon propose click tuning for the UHF channels. Perhaps when these problems are solved advertisers will purchase at remunerative rates all of the air time allocated by the FCC to commercial television. But I am skeptical, because of the nature of the market in which air time is at present sold.

Air time is sold to advertisers and what they are buying is not entertainment but an audience for their advertising. Because television is one of many advertising media and advertising only one of several business methods for promoting the sale of a product, the amount of money that an advertiser is willing to pay for broadcast time is limited. He will in fact pay only a few pennies per viewer. This means that a costly program is not feasible unless it attracts a very large audience. Therefore, multiplying the number of channels, as by licensing additional UHF stations, must soon cease to be commercially practicable as long as advertising remains the sole support of commercial television. The more channels there are, the smaller will be the audience for each and the less likely it will be that advertising revenues can cover the programming costs of all of the channels. From this it follows that cable television, insofar as it merely makes available a very large number of channels for television, is unlikely to change markedly the character and amount of programming. However, cable television does more: it facilitates the introduction of pay television, which would end the total dependence of television on advertising revenues.

One reason why the prospects of pay television on the air have remained dim is its high cost. Subscribers to Hartford's experimental pay television station were required to pay a $39 annual decoder rental to cover the cost of scrambling and unscrambling the broadcast signal (without scrambling, nonsubscribers would have received the signal free of charge), and this for only one channel of pay television.[12] Moreover, the charge to the subscriber for individual programs included substantial billing, accounting, and other operating expenses exclusive of program costs—which were in fact only about 15 percent of total costs. The nonprogram cost component would be substantially lower were pay television provided on channels of a cable television system. The cable operator who wanted to experiment with pay television would not have to buy or lease an entire broadcasting station, but could use one or more channels in his existing system. A simple filter to deny a channel to a cable subscriber who had not paid to receive that channel would replace expensive scrambling and unscrambling equipment. Billing and accounting costs would be lower because the system must bill the subscriber anyway for the basic monthly fee.

---

[11] See [16], pp. 27, 44.

[12] See Subscription Television Service [39].

If the demand for pay television is reasonably elastic, and if the marginal cost of providing pay television on the cable is substantially less than the cost of providing the same service over the air, then we can expect the supply of, and demand for, pay television, on the cable, to be much greater than McGowan, Noll, and Peck's very modest projection of the Hartford experiment would indicate.[13] There are two additional points: Multiple channels of pay television, fully practicable on the cable, would permit a better matching of program types to viewers' preferences; and if the same pay television programs were carried by many cable systems, program costs could be spread over many more viewers than were within the broadcast radius of the Hartford station.

An important factor is the attitude of regulatory authorities toward the provision of pay television by cable operators. The development of an over-the-air pay television service has been crippled by regulatory limitations. Even the FCC's newly liberalized rules authorize the establishment of a pay television station only in markets that have at least four other broadcasting stations, and strictly limit the sports events, motion pictures, and series-type programs that a pay television station may carry.[14]

The program ("siphoning") restrictions have been applied to cable operators wishing to institute pay television,[15] but no other restrictions have been imposed and in general the FCC seems less worried about pay television on the cable. It may feel that since cable television is a pay service to begin with, a prohibition against a charge for additional service would defy rationalization. Perhaps it recognizes that since the cable industry has a dual interest—in program revenues and in revenues from the basic subscription fee—it is unlikely to emasculate free service by bidding away all of the popular programs from advertisers for direct sale to viewers: it would lose many marginal subscribers. Whatever the precise reason, pay television on the cable may well encounter less regulatory obstruction than over-the-air pay television. Cable television may therefore provide the decisive opportunity for the establishment of a viable system of pay television.

This observation should be welcomed by those who believe that present television fare lacks richness and diversity. The essential fact is that the advertiser is not an adequate proxy for the viewer. His basic objective is to maximize audience coverage. The aversion of advertisers to supporting programs that are offensive to important groups in the community indicates that viewers will sometimes retaliate against an advertiser; perhaps there is a complementary tendency of viewers to reward advertisers whose programs please them greatly. But these tendencies are weak. They are in conflict with the individual viewer's self-interest, which is best served by separate selection of product and program. The viewer knows that his decision to buy one more (or less) box of cornflakes than he actually wants will not affect his television options.

---

[13] See McGowan, Noll, and Peck [16], pp. 29–32. They recognize that their projection may be too low. *Ibid.*, p. 32.

[14] See Fourth Report and Order (Subscription Television) [33].

[15] See Memorandum Opinion and Order in Docket 18397 (CATV) [35].

Under a system of pay television, small audiences having strong preferences for particular types of programs would sometimes bid away time from large audiences having weak preferences for other types. This does not happen today. Even if an advertiser could measure the intensity of viewing preferences, he would not sponsor a program preferred by a smaller audience unless the viewers' program preferences fortuitously accorded with their product preferences. Direct viewer payments, moreover, might greatly increase the resources invested in television programming. Advertisers have better substitutes for television as an advertising medium than do viewers for television as a medium of entertainment and information. One can think of many programs for which viewers would probably pay much more than a few pennies and which are not broadcast today only because they cannot command a huge audience. An audience of 200,000 willing to spend $2 each can pay for an hour of expensive programming; an advertiser would insist on an audience 100 times (or more) as great.

By making the fractionation of the total audience for television programs economically feasible, pay television would create a market for many channels, which are just what cable television can provide. The opportunity costs of providing many channels over the air would, in contrast, be extremely high because of the need to preempt spectrum space now used by other services.

Since the growth of pay television on the cable would be encouraged by repealing the "siphoning" restrictions that the FCC has imposed, we must consider whether those restrictions are justifiable. The concern that underlies them is that unregulated pay television (whether carried over the air or by cable systems) would bid away the most popular programming from the "free" broadcast service and thereby force people (some of them poor) who now watch these programs without charge to pay for them. An inadequate answer is that "free" television isn't really free because consumers of goods and services advertised on television pay for television programs in the form of higher prices for advertised products. It is not obvious, however, that such products would be cheaper if they were not advertised on television. Perhaps they would be more expensive because substitute methods of advertising or sales promotion were less efficient; or perhaps the consumer would spend more money on finding the right product. Even if not free, advertiser-supported television may be cheaper to the viewer who is a consumer of brand-name products than a system of pay television would be. Viewers might pay more for television than advertisers and might pay on top of advertisers, as in the newspaper and magazine industries, in which case it would appear to be inescapable that pay television had resulted in raising the price of television to the public.

This conclusion ignores, however, the incentive of program producers to devise substitutes for any program bought by a pay television station or network. These substitutes would be offered to advertiser-supported stations; and the pay television company would find it difficult to charge a price for seeing the original program when a substitute program was available to the viewer without charge. This of course assumes the continuation of free television, an important qualification to which we shall return.

For some television programs there may be no adequate substitute. But it is easy to exaggerate the unique and nonduplicatable in television programming. If, for example, the television revenues of professional football teams are very great, there will be more teams and greater promotion of sports such as soccer that have similar viewing characteristics. Every television viewer knows that successful formats in television entertainment are replicated endlessly.

What is likely is not that pay television will result in viewers' paying to see the identical fare that they now see for nothing, but that it will enable them to purchase television programs that they cannot obtain now on any terms (such as the opera that costs $400,000 to produce, has a potential nationwide television audience of only one million, and is not broadcast because advertisers will not pay $0.40 per viewer for television advertising time). Stated otherwise, pay television will enable viewers to substitute television programs for other goods that they value less. That is why the motion picture exhibitors are so hostile to pay television. They fear that people would rather pay $1 to see a first-run motion picture in their home, without advertising, than pay $3 to see it in a movie theatre.

The restrictions imposed by the FCC on pay television, both on the cable and over the air, may appear to let us have our cake and eat it too. They prevent the siphoning of popular television programming but leave pay television free to offer the type of program that is not carried on free television because the audience is too small. That is theory: practically, the restrictions may kill pay television in its infancy. They will certainly retard its growth.

Rather than restricting the growth of pay television, the goal of policy should be to assure competition in television distribution. As long as advertisers are able to obtain programs at cost, which are then shown to the viewer without charge, the viewer-consumer is protected, with respect to the wide range of programs that are not unique, against being forced to pay exorbitant charges. Fortunately, even the rosiest predictions of cable penetration indicate that the demise of over-the-air broadcasting is remote.[16]

☐ **Monopoly regulations.** The attractive possibilities for pay television on the cable should reduce concern that cable television, as a medium of diverse and high-quality programming, is likely to founder on the same shoals as over-the-air television. But perhaps the new medium eliminates one set of programming problems only to create another—monopoly control over programming.[17]

In general, neither economic theory nor empirical study supports the belief that an unregulated monopolist will reduce the quality of his product in order to increase his profits.[18] He will provide whatever level of quality the consumer is willing to pay for. Faced with a schedule of monopoly prices for alternative levels of quality, the consumer may choose a lower (cheaper) quality level than if competitive prices were charged, but the evil in such a case is not product degradation; it is monopoly pricing. However, when the monopolist's rates are regulated—a likely outcome in the cable television industry—the analysis becomes more complicated. If price alone and not program quality were regulated, the cable monopolist would have a strong incentive to reduce the quality and hence cost of his programs in order to recapture some of the monopoly profits eliminated by the price control. The regulatory agency would try to follow him down, but it is difficult to measure a regulated firm's costs, and frequent changes in those costs due to altered quality complicate the regulatory task.

---

[16] We return to this issue in Section 5.

[17] Assuming, perhaps prematurely, that cable companies will be effective monopolists in their service areas—on which see Section 4. There we also consider the problem of cable monopoly insofar as it is assumed to lead to higher subscriber rates rather than to less program choice.

[18] See Posner [24], pp. 584-85, for a review of this question.

Even if the refinement just introduced is disregarded, the previous analysis was too simple; it ignored the element of nonpecuniary profit. A cable monopolist doesn't have to take all of his profits in the form of money; he can take some in the form of the satisfaction or sense of power that some people derive from preaching to others. In providing a program mixture different from what his subscribers prefer, he will not be able to charge so high a price, but the reduction in pecuniary profits may be smaller than the increase in nonpecuniary satisfaction. Regulation may encourage the substitution of nonpecuniary for pecuniary satisfactions if, as seems reasonable to assume, only pecuniary profits are effectively constrained by regulation.[19]

This is not strictly a phenomenon of monopoly or even of the communications industry. A farmer can scratch his initials on every radish that he sells, and while his costs will be higher and his net money income lower than that of other radish farmers, as long as the satisfaction he obtains exceeds the reduction in his pecuniary income he is acting rationally; the communications industry is simply a more plausible field for such endeavors. Monopoly makes the consequences of a seller's indulging his preference for nonpecuniary profits more serious than they would be in a competitive setting. If there are 30 media outlets in a market and one is owned by a crank determined to propagate his opinions, the effect on the viewing audience will be minimal because they have many alternatives. Even if there are several cranks, there is no cause for alarm because they are likely to be cranky in different ways, so that the public will be exposed to conflicting points of view. Problems of a different order of magnitude are presented if the same crank controls all of the media outlets in a market.

The propensity to trade pecuniary profits for the nonpecuniary satisfactions of seeking to mold public opinion is probably related—inversely—to the absolute size of the enterprise. A large company is less likely to be owned by an individual or family than a small one, and where ownership is widely dispersed the managers of a company will find it difficult to identify public issues on which a partisan stand would not offend large numbers of shareholders. On reflection it is not surprising that charges of unfairness in broadcasting have been directed primarily against the owners of small radio stations, independent television stations, and family-owned communications companies.

---

[19] See Alchian and Kessel [1].

One partially offsetting factor should be noted. The larger an enterprise, the more likely that some of its activities will themselves be newsworthy, and these activities may not be fully and fairly reported by a media outlet owned by the enterprise. This was a major consideration that led the Department of Justice to oppose the merger of ABC and ITT before the Federal Communications Commission.[20] The Department pointed out that ITT's position as a major defense contractor and owner of telephone companies and other telecommunications interests in Latin America might prejudice ABC's reporting of defense and Latin-American affairs.

What are the implications of this analysis for cable television? Much of the programming carried by cable companies is beyond the power of the companies to alter or influence, for it consists of the programs of television stations whose signals are carried on the cable. Moreover, even with general rate regulation, the rates charged subscribers for special channels (or programs) are likely to remain unregulated, so that cable systems will retain a strong incentive to provide optimum service at least on those channels. Degradation of quality does not seem a serious danger.

The danger of local monopolies of the sources of opinion may well diminish rather than increase as cable television grows. By increasing the number of signals available in the home, cable television multiplies the subscriber's independent sources of news and opinion. This is an especially important factor in smaller markets, where in the absence of cable television households would receive only one or two signals; and it is in the smaller markets that the FCC has been most liberal about allowing cable systems to carry distant signals. Insofar as a cable system also originates programs on one or more channels, it creates an additional independent source (or sources) of news and opinion in the market.

Yet it is precisely the ability of a cable operator to originate on more than one channel that has led to proposals to restrict his control of channels on which he originates.[21] The cable operator who originates on two channels has been likened to a company that owns two television stations in the same market, a type of ownership forbidden by the FCC's misnamed "duopoly" rule.[22] The analogy is strained. The owner of a television station who purchases another such station in the same market thereby reduces the number of independent sources of news and opinion by one; the cable company that decides to originate on a second channel does not thereby alter the number of independent sources at all. Furthermore, the likelihood of multiple origination by the cable operator may be exaggerated. Programming and distribution on the cable are distinct functions. Distribution involves the installation, maintenance, and replacement of a grid of wires. The cable operator carries out distribution, and he also decides which broadcast signals in addition to those of the local stations (which he is required by FCC regulation to carry) to put on his cable—a programming function. If on top of this he procures the particular programs to be shown on one or more of "his own" channels, his operation becomes highly complex.

---

[20] The story is told in Green [10], Vol. 2, pp. 853–59.

[21] See, e.g., Notice of Proposed Rulemaking and Notice of Inquiry in Docket No. 18397 (CATV) [36].

[22] This rule is described in Barnett [2], p. 251.

In these circumstances one would expect cable companies, without any prodding from government, to lease channels to firms specializing in the programming function. As the cable industry grows, we can expect the tendency toward vertical disintegration to grow too: the combination of functions observable in today's cable systems is a sign of the early stage of the industry's development.[23]

But leasing may not be a complete solution. As long as the cable operator is free to pick and choose among prospective lessees he may reject those whose programming policies do not accord with his preferences. A superficially attractive response is to forbid cable systems to discriminate among lessees of their channels. The difficulty is in defining and administering an adequate standard of nondiscrimination. The enforcing authority would perforce become deeply involved in the methods used by cable systems to price their channels and therefore in their methods of cost accounting. It would have to consider such thorny questions as whether a cable operator was justified in charging an additional price to a lessee whose unpopularity (for whatever reason) with a large segment of the viewing public was likely to impair the cable system's goodwill and revenues.

Either compulsory leasing or a rule forbidding discrimination among lessees seems too drastic, too far-reaching and, in light of experience with similar controls in the public utility and common carrier industries, too easily abused a measure to be imposed when it is as yet unclear what patterns of origination or leasing will emerge as the industry matures. Such measures might become irresistible, however, if cable systems were to affiliate with television stations, networks, newspapers, and other sources of news and opinion on a large scale. Where the only television station, newspaper, and cable system in a community are owned in common, one can still take comfort in the fact that the cable system will be importing the signals of stations not affiliated with the system; but it is small comfort, especially with regard to local news and events, in which distant stations presumably are uninterested. It is probably easier as an administrative matter to prevent concentration of a community's mass media in one set of hands than to accept the concentration and then try, through common carrier regulation, to force the monopolist to grant access to other voices on reasonable terms.

This consideration provides a practical reason for continuing (and perhaps extending to newspapers) the FCC's prohibition against common ownership of cable systems and television stations located in the same market,[24] while forgoing, at least for the present, rules governing leasing of channels. To be sure, a likely motive for the acquisition of cable systems by owners of competing media is a desire, akin to that which motivated the diversification program of the cigarette manufacturers, to move gracefully out of a declining industry. This process should be encouraged in the present context as in that of the cigarette industry in order to reduce the pressure exerted by the industry for protection against the causes of its decline. Furthermore, cross-media ownership, by increasing the size of cable enterprises, reduces, according to the earlier analysis, the likelihood that systematic influence molding will be attempted. On balance, even the case for ownership regulations is only marginally persuasive, albeit restrictions on cross-media ownership in a single market—which is all that I tentatively approve—would leave room for growth and diversification through media acquisitions across markets.

---

[23] Cf. Stigler [28], pp. 135–36.

[24] See Second Report and Order in Docket 18397 (CATV) [38].

4. Regulation of subscriber rates

■ **Nature and magnitude of the monopoly problem.** A cable system's grid or network of cables is somewhat akin to that of the local water, electrical, gas, or telephone company. As with the supply of these other services, running more than one company's cable to any home would involve unnecessary duplication, for a single cable can carry all of the signals that a subscriber is likely to want and at substantially lower cost than two or more smaller cables. These characteristics raise the possibility of monopoly pricing, which leads to a reduction in output below efficient levels and, as mentioned earlier, would reduce the scale and growth of the cable television industry. However, some important qualifications must be borne in mind in the present context. To begin, the point that monopoly reduces output is strictly valid only if the monopolist is constrained to sell at a single price. Ordinarily, the assumption is justified, since if a seller sold at different prices to different purchasers those who bought at lower prices would resell to those charged the higher prices, and the monopolist's attempt to discriminate would fail. For sellers of nontransferable services such as cable television, however, arbitrage among purchasers is not much of a threat, and discrimination becomes a feasible strategy. A cable company would find it easy to discriminate. It could charge different rates in low- and high-income areas, a basic subscription fee and then separate charges for additional channels and/or programs, hourly rates, and so on.

Under perfect discrimination the monopolist's output would be identical to that of a competitively organized market. The only difference would be in the transfer of wealth from consumers to the monopolist. However, perfect discrimination is unattainable in the real world, and when less than perfect, its effect on allocation is indeterminate; in some cases it may be worse than if the monopolist charged a single price.[25] Price discrimination may also entail substantial administrative costs. Nonetheless, the existence of excellent opportunities for discrimination in the provision of cable service reduces one's confidence that unregulated cable monopolies would result in a substantially suboptimal output.

If the over-the-air service provided by local television stations is a good substitute for cable television service, cable companies will not be able to obtain excess profits. That some consumers are willing to pay for the better reception (especially of color) and greater number of signals that cable television offers is an inconclusive indication of monopoly power; the price that consumers are willing to pay for cable service may just cover the costs of the service.

No problem of monopoly need arise, furthermore, if purchasers have an adequate opportunity to solicit competing bids in advance. The winning bidder will be the supplier who undertakes to provide

---

[25] See Robinson [27], pp. 190–95.

the service at the lowest possible price, a price that will not include any monopoly toll.[26] As the cable television business now operates, subscribers are rarely if ever given a choice between cable companies; only one cable company solicits their patronage. The immediate cause of this, however, is not any inherent characteristic of cable television but the fact that a cable company must obtain a municipal franchise in order to be permitted to serve any part of the community. Whether it is because they assume that the cable television business is a natural monopoly, and they desire to limit the (surely minor) inconvenience to the public of having several companies using public rights of way to string or lay cable, or alternatively, because they seek a share of monopoly profits in the form of franchise fees, municipalities do not grant more than one cable franchise in any area within their jurisdiction.

Finally, even if cable operators enjoyed effective monopoly power, it would not follow that they would exercise it by raising subscriber rates above cost. They might prefer to charge very low or even zero rates in order to maximize coverage, and look to program fees or fees from networks, advertisers, or channel lessees—or some combination of such charges—as the source of their profits.[27] That would not eliminate the problem of monopoly, but it would eliminate the need for regulating subscriber rates.

The foregoing discussion may seem a bit unreal in light of widespread opinion, some of it supported by evidence, that cable operators enjoy astronomical profits.[28] Assuming there are such profits, we do not know to what extent they are due to technical monopoly and to what extent to municipal franchise policies that foster monopoly unnecessarily.

☐ **Nonregulation.** The last point raises an intriguing possibility in dealing with the problem of monopoly rates to subscribers: the elimination of all franchise regulation. To be sure, were there no franchise or other public regulation of the transactions between cable television companies and subscribers, and if only one company solicited subscribers in an area, the subscribers might be induced to pay a fee in excess of the cost of serving them. However, in the absence of regulation, more than one company might solicit in each area, since many companies are capable of building and operating cable television systems. In principle, competition among the soliciting companies would squeeze any monopoly profit out of the subscriber charge, as long as the companies were prevented from colluding.[29]

There are admittedly many problems. For one, if during the period of initial competition several companies succeeded in signing up some of the residents of an area, each company's subscribers might be widely scattered. The building of a separate cable network by each company in such circumstances would involve unnecessary duplication of facilities. However, one would expect the companies to exchange subscribers in advance of construction so that each would be serving a compact group. The subscriber who found himself in the service area of a company other than the one to whose service he had subscribed would be protected by his contract with the first company. If the second company wanted to provide a different level of service from that specified in the subscriber's contract with the first, it would have to obtain the subscriber's consent.[30]

---

[26] See Demsetz [7].

[27] Alternative methods of pricing cable service are discussed in Ohls [20], pp. 452–57.

[28] Such as the McGowan, Noll, and Peck study [16], discussed in Section 2.

[29] The result would be like that under a "full-requirements contract," where the purchaser agrees to buy all his requirements of a particular good from one supplier, at a specified price and for a specified period of time. During the life of the contract, there is only one seller; he has a legal monopoly. But the buyer is protected by the initial competition among potential sellers seeking to be awarded the exclusive contract.

The contract solution is not free from difficulty. When the cable subscriber's contract with the cable company expired he would be at the company's mercy, for there would be no other cable company in the area to which he could turn. To be sure, the monopoly position in which the cable company found itself upon expiration of its contracts with subscribers would persist only until a competing company could move into the area and offer them a better contract. However, there would be a lag between the time the first company raised its price and a new company found out, signed up enough subscribers to justify building its own system, and either built such a system or bought the existing system of the first company.

The size of the area in which only one company can economically provide cable television service is probably the most important factor in determining the length of the lag. If that area is an entire city, it may take a long time for a new entrant to win over the existing cable system's subscribers and build a cable network to serve them. If the area is only a small neighborhood, entry could be rapid; cable companies in adjacent areas would only have to extend their lines a few blocks.

One might argue that new entry would be unlikely until the plant of the existing cable company was no longer usable and had to be replaced. Since the costs of operating an already constructed cable system are only a small proportion of the total costs of the system (most of the costs being incurred in the initial construction), it might appear that the existing cable company in an area would have a decisive cost advantage over any new firm which would have to construct (or extend) its system in order to serve the same area. But such reasoning again ignores the possibility of contractual arrangements. Before entering a new area, a cable company can be expected to seek contracts from subscribers now served by the firm (or firms) already in the area. Such a contract might provide that the subscriber, in exchange for a lower rate, would agree to take service from the new firm for a specified period. During this period, the firm would be protected by its subscriber contracts from retaliatory price cutting by the existing firms. To be sure, rather than lose this business the existing firms might offer subscribers a lower rate in order to forestall new entry. But that is all to the good. The process of threat and response would eventually compress rates to a competitive level.

Although rapidity of entry would obviously improve the subscriber's bargaining position, cable subscribers might not necessarily be subjected to monopoly prices if entry were slow. The feasibility of contractural arrangements is again critical. Suppose that it it would take two years for a new company to enter the market of a company that has three-year cable service contracts with its subscribers. At the end of the first year of the contract, subscribers should be able to obtain either an extension of the contract or a new contract with a rival cable company for service commencing with the expiration of the original term. The subscriber has the upper hand.[31]

---

[30] Companies might not offer contracts—that is, enforceable commitments—at first, but competition would soon force them to offer such contracts were it common practice for cable companies to jack up rates once the system was built.

But we have ignored the problem of contract expense. Most transactions involving individual consumers are not handled by means of formal contracts embodying continuing obligations, probably because of the cost of such arrangements (warranties, however, are an important exception). A contractual solution to the problem of cable monopoly could turn out to be quite costly; whether it would or not is an unstudied empirical question.

☐ **Franchise regulation.** The legal authority of municipalities to refuse the use of public rights of way can be used to affect a cable company's ability to exploit its monopolistic position (assuming it has one) in one of two basic ways. The first, which I shall call the "concession" approach, and which has already been employed widely, can take a number of forms. The simplest is to auction off the franchise to the bidder who offers to pay the largest lump sum for it. In bidding under such a system, an applicant will first estimate the monopoly profits (if any) that the franchise would yield were no fee imposed. He will then estimate the capitalized present value of those anticipated profits. That amount is the maximum amount that he will bid for the franchise. The bidder who expects the franchise to yield the largest monopoly profits will be the high bidder and receive the franchise.

This system captures the monopoly profits for the public (in the form of the lump-sum payment to the franchising authority) and is easy and cheap to administer, but it does nothing to alleviate the monopoly problem and may actually aggravate it. Fees to subscribers are no lower than if no attempt to regulate the cable monopoly were made. Less obviously, there is a danger that the franchising authority will be tempted to offer applicants not merely the right to wire homes in its jurisdiction but the exclusive right to do so. By adding a legal monopoly to the franchisee's natural monopoly, the franchising authority may be able to extract a larger sum from the franchisee, but at considerable social cost; for the grant of long-term exclusive franchises to cable companies may retard the process by which changes in technology can, over time, erode a natural monopoly.

An alternative approach would involve awarding the franchise to the firm that offered to pay the franchising authority the highest fee per subscriber or the largest percentage of gross revenues. This method is dubious because it could lead to even higher subscriber fees and lower output levels than the lump-sum auction method. If the fee is based on the number of subscribers, the franchisee will treat it as an additional cost per subscriber. The profit-maximizing strategy in these circumstances is to raise rates to subscribers.[32] If the payment to the franchising authority is a percentage of gross revenue, the cable operator will treat it as reducing the demand for his product; and again one can demonstrate that as long as the cable operator seeks to maximize profits he will raise his subscriber rates. Thus we end up with subscriber rates even higher than those of either an unregulated monopolist or a monopolist franchised under the lump-sum auction method outlined previously.[33]

---

[31] I do not assume that the subscribers will bargain collectively with the company that serves them. It is only necessary that there be potential entrants to whom individual subscribers can turn for a better deal.

[32] For an explanation of why this is so, see Posner [23], p. 16.

Another alternative would be to award the franchise to the firm that offered the most service—for example, the firm that promised to dedicate the greatest number of channels, free of charge, to the use of the franchising authority, or to provide the greatest amount of original programming. The effects on output of this approach are more complicated but still adverse. If the response of applicants under such a scheme is to offer more channels or more programs, the consequences depend on the additional investment and operating cost entailed in the larger system. That cost appears to be substantial[34] and will lead to a higher price to the subscriber and to fewer subscribers being served.

The franchise applicant under pressure to offer service concessions may not respond by proposing a larger system. To illustrate, suppose he estimates that subscriber demand would warrant an eight-channel system and that a winning bid would require that he dedicate two channels to the franchising authority. He may offer to build a nine-channel rather than a ten-channel system. Since the subscriber demand for seven channels is less than the demand for eight, he will attract fewer subscribers and hence less revenue, but if the smaller system is sufficiently less expensive to build it may represent, on balance, the better bid.

The usual practice in awarding cable franchises is different from the possibilities we have been discussing. Rather than allowing the size of the concession to be determined by competitive bidding, franchising authorities typically specify a fixed percentage of gross revenues, or other specific concessions, to be provided by the franchisee. This procedure limits the undesirable effects of encouraging franchise applicants to outbid each other, but it does not eliminate them. Any franchise exaction that increases the per-subscriber cost to the cable operator will be passed on, in part at least, to the subscriber in the form of higher rates. Moreover, the fixed concession method deprives the franchise system of one of its most attractive features, ease of administration, since when the franchise is awarded to the high bidder the process of choosing among applicants involves a minimum of administrative machinery and official discretion.

The franchising power could be employed quite differently: not to compel the cable monopolist to divide monopoly profits, either in money or in service, with the franchising authority, but to induce him to charge subscribers the competitive rate. Under this approach a local agency would act as the purchasing agent of the residents in its jurisdiction for cable television service. It would conduct an auction, with the franchise awarded to the firm that offered the most attractive price-quality package from a subscriber standpoint. Each bidder would submit a plan of service and a schedule of rates. As long as there was more than one bidder and collusion among the bidders was prevented—conditions that ought not to be insuperably difficult to secure—the process of bidding subscriber rates down and quality of service up would eliminate monopoly pricing and profits.

---

[33] This method of awarding franchises may, indeed, be unworkable. Even when a bidder offers a fee per subscriber or a percentage of gross revenues so high that it would require him to operate at a price level where output was extremely small, he would still obtain monopoly profits. There is therefore room for a still higher bid involving a still smaller output. The tendency is for price to approach the highest point on the demand curve and for output to approach zero. The franchising authority may attempt to prevent such a result by fixing a ceiling on the rates that the franchisee may charge, but this involves a shift to another form of control, rate regulation, which has (as we shall see) severe problems of its own.

[34] See Comanor and Mitchell [6], pp. 170–72.

Since bidders may offer quite different types or levels of cable service, having different costs, it may be difficult to decide which bid is best from the subscriber's standpoint. Although consumers face and overcome this problem daily in choosing among products that differ in quality as well as price, here it would not be the consumers (the cable subscribers) who determined the preferred price-quality mix but an official body that must attempt to infer the consumers' preference and, failing that, to substitute its own. The problem could be alleviated to some extent by beginning the bargaining process with an "open season" in which all franchise applicants were free to solicit the area's residents for a set period of time. This would not be a poll; the applicants would seek to obtain actual commitments from potential subscribers. At the end of the solicitation period, the commitments received by the various applicants would be compared and the franchise awarded to the applicant whose guaranteed receipts, on the basis of subscriber commitments, were largest. In this fashion the vote of each subscriber would be weighted by his willingness to pay, and the winning applicant would be the one who, in free competition with the other applicants, was preferred by subscribers in the aggregate. To keep the solicitation process honest, each applicant would be required to contract in advance that, in the event he won, he would provide the level of service, and at the rate represented, in his solicitation drive.

In some instances this method would distort consumer preference. Suppose that 30 percent of the consumers "vote" for system *A*, 30 percent for system *B*, and 40 percent for system *C*, and that *A* and *B* are almost identical but *C* is quite different: *C* will be franchised, though franchising either *A* or *B* would please more consumers. Problems of this sort are resolved automatically in a market by recontracting. One can devise methods of approximating the market result—by having a run-off solicitation campaign or by permitting subscribers to sign up with more than one system—but these may be awkward and expensive.

Another problem of the bargaining method is the duration of the franchise. If it is long, the parties may not have foreseen all of the circumstances that might require modification of its terms. Although this is a problem common to all contracts, the peculiarity here is that one of the contracting parties is not a true party in interest but a public body charged with overseeing the interest of the other parties (the subscribers). Experience with regulatory agencies suggests that one cannot assume such a body will represent the consumer interest faithfully. When the cable company asks for a modification of the contract by virtue of an unforeseen change in circumstances, the public body may react ineffectually or perversely.

It is therefore desirable to keep the term of the franchise sufficiently short so that no modification of its terms need be entertained. However, the fact that the cable company's plant normally will outlast the period of its franchise raises a question: Will not the cable company be able to outbid any new applicant, who would have to build a plant from scratch? And will not the bargaining method therefore be ineffective after the first round? Not necessarily: in bidding for the franchise on the basis of new equipment costs, new applicants need not be at a significant disadvantage in relation to the incumbent franchisee. For example, once a new applicant is franchised he could negotiate to purchase the system of the existing franchisee, who faces the loss of the unamortized portion of his investment if his successor builds a new system. Insofar as the economic life of cable plant is considered a problem when the franchise term is short, it can be solved by including in the franchise a provision requiring the franchisee, at his successor's option, to sell his plant (including improvements) to the latter at its original cost, as depreciated.

☐ **Rate regulation.** Franchise regulation (or deregulation) is only one possible approach to the problem—if there is a problem—of monopolistic subscription rates. Another is direct regulation of those rates. While there has been as yet little direct experience with attempts to regulate cable rates, there has been a good deal of experience with rate regulation in other industries, such as local gas, water, electrical, and telephone service, that resemble cable television in their supply and demand characteristics. That experience offers some guide to the likely costs and benefits of extending the principles of public utility regulation to cable television.

In principle, placing a ceiling on the rates that a monopolist may charge is a straightforward and efficacious corrective. But there is reason for concern that in practice it (1) may be ineffectual and (2) may impose substantial, if largely invisible, efficiency costs on the community.

Thus far, empirical studies have not discovered any appreciable impact of regulation on the rates of the companies regulated.[35] That is probably due in part to the paucity of resources, especially at the state level, devoted to rate regulation, and in part to the difficulty of determining a regulated firm's costs—a necessary step if its rates are to be set at a level that will limit or prevent monopoly pricing while permitting the firm to recover its legitimate expenses. Some costs can be determined easily enough if the underlying data are known, but such data are held by the regulated firms and are not always easy to extract in a trustworthy or usable form. Other costs are difficult to determine even if the underlying data are easily available. These include depreciation and the cost of equity capital. The latter depends on the rate of return that investors demand for providing capital to firms of the regulated company's risk class, and on the value of the company's capital assets against which to apply the rate; both quantities as yet elude confident measure. Problems are also encountered in separating costs between regulated and non-regulated services provided by the company. Slight errors in the successive steps required to determine the company's relevant costs may cumulate so as to deprive regulation of any significant effect on the regulated firms' rate and profits. Furthermore, rates are reviewed periodically rather than continuously, and changing conditions of cost or demand may produce windfall profits or losses in the intervals between rate reviews.

---

[35] These studies are discussed in Posner [25]. A study not discussed there is Moore [17]. For a fuller treatment of the matters discussed in this section of the paper, refer to Posner [24].

None of these problems can be written off as trivial in the cable television context. On the contrary, with the industry in so embryonic a state of development, and demand and technology still so fluid, the determination of proper depreciation periods and of risk (and hence of the required rate of return to equity financing) is likely to be unusually difficult; periodic rate reviews may be too infrequent to permit the adjustment of rates to rapidly changing costs and demands; and since many cable companies have systems in different jurisdictions, and since some companies are subsidiaries or affiliates of non-cable enterprises, cost separation problems seem unavoidable. While present cable systems are relatively small and simple compared to many other regulated firms, the picture may change dramatically as the industry matures. And even if the difficulties in the way of effective regulation prove less pronounced with respect to cable television than with respect to some other industries, the resources devoted to its regulation are likely to be less ample as well. Legislatures often thrust additional duties on regulatory agencies faster than they appropriate additional funds, believing that the agencies can probably do more with their existing resources than they are doing.

The costs of regulation are commonly measured by the budgets of the regulatory agencies; but this procedure greatly understates the true costs. It ignores not only the legal and other expenses that are imposed upon regulated firms, which exceed by several times the expenses of the agencies,[36] but also the considerable indirect costs of regulation arising from (1) the adaptive responses that firms will make to a regulatory environment and (2) the perverse applications of regulation that are such a common and seemingly inescapable feature of public utility and common carrier regulatory schemes. This is not the place to review the depressing literature on the costs of regulation. Suffice it to say that neither in general nor in the particular circumstances of the cable television industry can it be complacently assumed that public utility controls would have a positive benefit-cost ratio.[37]

We are left, then, with a set of alternative approaches to the regulation of cable monopoly, no one of which is free from substantial difficulties. A period of experimentation would seem to be indicated.

5. Protection of broadcast television

■ The original thrust of federal regulation of cable television was to protect over-the-air broadcasters from competitive injury at the hands of the new industry. That thrust is losing momentum, in part because the prospective injury to over-the-air broadcasters, and especially to the UHF stations that are a particular object of the FCC's solicitude, has been shown to be less severe, or at least less immediate, than originally believed;[38] but it retains a good deal of force.

---

[36] A case study is Gerwig [9].

[37] See Posner [24, 25].

☐ **The threat to broadcast television.** The subsequent subsections assume that cable television, unless restricted by the government, will make substantial inroads into the amount and quality of over-the-air television service. The assumption is probably unsound. Even if cable television were to become dominant, it does not follow that over-the-air service would decrease. If 90 percent of the country's households subscribed to cable television, the remaining 10 percent would still constitute an enormous market for advertisers. And many cable subscribers would own portable television sets. Since programming, rather than transmission, accounts for the major costs of television, the self-interest of advertisers would in all probability lead them to defray the modest costs of operating a sufficient number of transmitters to reach households not on the cable and owners of portable sets.

The principal change would be in the number of stations. The reason for having 800-odd television stations is not to maximize the amount and quality of television received in each household—a much smaller number of regional stations would permit a larger number of signals to be received by each viewer—but to preserve the curious concept of a nation of local television stations.[39] High cable penetration would make many of these local stations uneconomical and thereby foster a long overdue reorganization of the industry into a much smaller number of regional stations. Such a reorganization would confer additional social benefits by freeing up large portions of the radio spectrum for other uses.

It is sometimes argued that cable systems would buy exclusive rights to programs (and thereby deny them to the over-the-air service) in order to attract new subscribers with entertainment not available over the air. But whether this promotional device would be used extensively is conjectural. A cable system would have to pay a higher price for exclusive than for nonexclusive rights to a program, since in the former case the program owner would lose the revenues that he would have obtained by selling rights to television stations as well as to cable systems. A cable exclusive may injure broadcast competitors by making their service less attractive to viewers but it must also injure the cable system by requiring it to pay more for programs.

☐ **Arguments for protection.** Let us assume, contrary to the previous discussion, that cable television (quite apart from any pay-by-the-program feature) would lead to a reduction in the quantity and quality of over-the-air service; and let us consider, to begin, several respects in which cable television might be thought to have an unfair advantage in competition with broadcast stations as a result not of intrinsic merit but simply of unequal treatment in the eyes of the law. First, broadcast stations must pay royalties when the matter they broadcast is copyrighted, while cable systems that rebroadcast those signals need not. That is the effect of the present Copyright Act as interpreted by the Supreme Court in the Fortnightly case.[40] Second, broadcast stations are expressly forbidden by Sec. 325 of the Communications Act of 1934[41] to rebroadcast any other station's signals without that station's consent; no similar provision is applicable to cable systems. Third, it is FCC policy to discourage broadcast stations from rebroadcasting another station's signals regardless of consent. A station that violated this policy would risk nonrenewal of its license.[42]

---

[38] See Park [22].

[39] Discussed in Johnson [11], pp. 2–3.

The copyright argument seems weak. A broadcast station is in effect a middleman between the owners of programs (copyright holders) and advertisers. The station buys popular programs in order to generate audience and sells time in breaks between and within programs to advertisers wanting to reach a large audience. Revenue from advertisers covers the royalties to the copyright holders. A more popular program will attract a larger audience and enable the station to charge higher rates to advertisers, and will also command a higher royalty. By carrying a station's signal beyond its ordinary range, a cable system enlarges the station's audience and thereby increases the revenue that the station is able to extract from advertisers. This in turn enables the copyright holder to extract a larger royalty. At least as a first approximation, then, both station and copyright holder benefit from cable service even though the cable system pays no royalty. Indeed, it seems anomalous to require it to pay royalties when it is not in the business of selling audience to advertisers.[43]

Since television stations located in the cable system's market must pay royalties for its programs, the question arises whether those stations will be harmed as a result of the ability of the cable system to obtain programs free of charge, even if the distant originating station and the copyright owner are adequately compensated. The local station need not suffer any harm. It is free to sell time to advertisers on the program that it buys; the cable system cannot do so unless it likewise buys the programs. What would be unfair to the local stations would be to impose more onerous conditions on them when *they* desire to carry distant signals. Here we come up against Sec. 325. If cable systems are not required to obtain the consent of the originating station, this provision should be repealed.

---

[40] Fortnightly Corp. v. United Artists Television, Inc., 392 U. S. 390 (1968).

[41] See [29] at 325.

[42] There is a limited exception for "translators" that extend television coverage to remote communities. See 47 C.F.R. 74.731 *et seq.*

[43] If cable systems blocked out the commercials on the station signals and sold the time to other advertisers without paying royalties to the copyright owners, they would be guilty of piracy. The cable system would be selling program time to advertisers, while refusing to buy the programs from their owners. However, instances of such substitution are apparently rare, and nothing in the Fortnightly decision suggests that cable systems could escape liability under the Copyright Act for such acts of piracy. The Court was careful to limit its decision to the situation in which the cable operator simply retransmits the signal of the distant station. In fact, the Court's holding of copyright immunity was expressly premised on the analogy of the cable system to a giant antenna erected by the set owner himself (see 392 U.S. at 399)—an analogy that collapses when the cable system does more than rebroadcast. The Court's ruling is also limited to the case in which the cable system does no original broadcasting (see *id.* at 392, note 6); inserting its own commercials would be a form of origination.

By the same token the FCC's policy against retransmission of broadcast signals by its broadcast licensees should be changed. One need not question here the Commission's preference, which appears to be at the root of the policy against retransmission, for local control of television programming. The policy against retransmission is not effective in implementing that preference. While forbidden to rebroadcast the signals of other stations, television stations can and do broadcast primarily network programs, network reruns, and other syndicated matter—none of which originates locally. What is the justification for a rule that permits broadcast stations to obtain programs from distant sources by electronic interconnection or mailed tape, and forbids them to obtain such programs only when the means of receipt is by interception of a broadcast signal?

Perhaps the answer is that a system under which both cable systems and broadcast stations were free to carry distant signals without copyright liability would not protect the legitimate interests of copyright holders. When a program is carried in a local market by a cable system, the owner of the program cannot charge the local stations in the market as much for the right to carry it, because he can no longer offer an exclusive contract. In some instances his loss will be made up by the larger royalty that he is able to extract from the originating station. But if the originating station sells time only to local advertisers who are indifferent to reaching an audience outside of the station's immediate market, carriage by the cable systems will not enhance the station's advertising revenue and enable it to pay a higher royalty.

However, this takes too static a view of the situation. A station whose signals are carried by cable systems into distant markets will market time to regional rather than local advertisers in order to capitalize on its larger, regional audience.[44] As such stations reach out for a regional market, local advertisers will switch to stations which lack regional aspirations and whose signals are not carried outside the markets in which they are located. Such a reassignment of customers would follow the pattern established in network broadcasting, whereby the network, which through its affiliates has access to national and regional audiences, sells time to national and regional advertisers while local stations sell time to local advertisers.

The more rapid the growth of the cable television industry, in fact, the better off copyright holders are likely to be. Those channels on which cable entrepreneurs plan to originate programs represent an important new market for copyrighted material. And the revenues that cable companies derive from subscriber fees represent an important potential source of augmented copyright royalties. Any rule that by limiting carriage of distant signals retards the expansion of cable television service will postpone these returns. How then to explain the opposition of the copyright interests to unrestricted carriage of distant signals? They may deem it wise strategy to oppose all attempts at limiting copyright liability. Some copyright holders, moreover, have broadcast interests. Most important, any legal hold that the copyright interests are able to establish over the cable companies will be worth something to those companies to remove.

---

[44] There is evidence that stations do advertise cable carriage of their broadcasts to attract additional advertisers. See Hearings on H.R. 4347 Before Subcommittee No. 3 [30], pp. 1281–82, 1302–06.

The foregoing analysis lends little support to the argument that the right of cable companies to import distant signals should be limited, although it does suggest that some relatively minor legislative and administrative modifications to permit broadcast stations freely to rebroadcast the signals of other stations should be adopted. But neither does the analysis support the opposite position: that a rule limiting the right of cable systems to carry distant signals must impede the growth of the service. Where transactions are permitted and are feasible, the initial assignment of legal liability will not affect resource use.[45] The qualifications in this statement must not be ignored: The parties must be free to bargain, and transactions must not be prohibitively costly. A rule under which permission to carry distant signals depends not on the willingness of the program owner or of his proxy, the originating station, but on the leave of the Commission violates the first condition. Were the FCC to forbid the copyright and cable interests to work out clearinghouse procedures (analogous to those employed in the recorded-music field) for the wholesale granting of copyright licenses, the second condition, reasonably low transaction costs, might also be violated; haggling over every single copyrighted item might be unduly costly.

Let us now consider what the impact on the existing pattern of income distribution in our society might be if cable television continues to expand. Were over-the-air service to be wholly displaced by cable, people in remote areas, where the cost of providing cable service was prohibitive, would be cut off from any television service. However, a policy of benefiting rural inhabitants in the lump would be undiscriminating; and a method of subsidizing the wants of rural dwellers that involved a denial or limitation of service to millions of other people would impose unnecessary costs on society. If it is believed that rural inhabitants should be assured television service, that object can be secured without limiting the availability of cable television to those who want and are willing to pay for it, either by paying television stations to remain in operation or by paying cable television companies to extend service to the rural population.[46] The point need not be labored further since we have already discussed the inappropriateness of internal subsidization as a goal in the regulation of cable television.

Another effect of the displacement of over-the-air by cable service would be to impose a charge for a service that viewers now receive gratis. This might be thought to work particular hardship on lower-income groups. However, all other considerations aside, limitations on cable service could well injure rather than benefit those groups. They may rather pay $5 a month for cable service, and $1 for occasional special programs, than receive over-the-air service for nothing. The superior reception, additional signals, and an occasional first-run movie thus obtained might make cable television an acceptable substitute for more costly entertainment, information, and educational services. It is not as if the poor were especially well served by the present system of free television service. The present system does not appear to cater to those wants or needs that are special to the poor. Geared to the tastes and interests of its predominantly middle-class audience, the television industry has offered little or no programming on child care, job opportunities, vocational training, and ethnic culture that might especially interest poor people. Accordingly, a policy of limiting the expansion of cable television service in order to assure that the poor continue to receive free television seems an impoverished, if not hypocritical, response to the problem of poverty.

---

[45] See Coase [4].

[46] The Rural Electrification Administration, with its programs of financial assistance to rural electrical and telephone cooperatives, provides a precedent for this approach.

There is one more ground on which to urge the protection of the broadcast industry against the competition of cable television. A firm that obtains a license from the FCC to operate a television station and succeeds in affiliating with one of the major networks will generally be able to derive a revenue from the sale of advertising time that is vastly in excess of its capital and operating costs. The government could capture these rents for itself by auctioning off broadcast licenses to the highest bidder. The amount bid would represent the bidder's estimate of the capitalized value of the anticipated rents. (When a licensee transfers his license to another, the purchase price includes just such an amount.) Instead, service in the "public interest" has been made the criterion for the initial grant and renewal of broadcast licenses.[47] In effect, the licensee is compelled to rebate a portion of his windfall profits to the public in the form of service beyond what an unfettered profit maximizer would charge. This portion may be very small, but at least some news and public-affairs programming is carried that is not remunerative. If cable television is permitted to compete away the large profits of the television stations, the public-interest programming supported by those profits may be curtailed. Because much of this programming involves the reporting of controversial political events, such as election campaigns and various governmental activities, government subsidy is not an attractive alternative.

As an argument for protecting the broadcast service from inroads by cable television, this overlooks the cost, in public-service programming foregone, that a policy of protection would impose. Cable television probably provides a better environment for the encouragement of public-interest programming than over-the-air service. With no lack of channels, such programming is not so apt to be displaced by more popular programs that can command greater advertiser support. And having subscriber as well as advertiser revenues, the cable system need not find a commercial sponsor in order to defray the cost of a program. If public-affairs and local-service programming is likely to flourish on cable television to a degree not attainable in over-the-air broadcasting, any policy of limiting the growth of cable television will curtail the public-interest programming available to cable subscribers, at the same time that it is perhaps preserving such programming for over-the-air viewers. Only if and when cable television becomes very widespread will the profits of network-owned and network-affiliated stations, the mainstays of the over-the-air service and the principal sources of the programming in question, be jeopardized; but at that point, by hypothesis, most viewers will be cable subscribers.

---

[47] See Communications Act of 1934, as amended [29]; Public Service Responsibility of Broadcast Licensees [37]; and Commission Policy on Programming [31].

My conclusion is that a policy of protecting broadcast television from the competition of cable service cannot be justified.

6. Rationalizing the structure of regulation

■ All three levels of government—federal, state, and municipal—have become deeply involved in the regulation of cable television. The thrust of regulation, however, has been markedly different at each level.

At the municipal level, where cable regulation has usually taken the form of conditioning permission to use public rights of way upon the franchisee's agreement to rebate a portion of his gross revenues and provide free service to the city, the thrust of cable regulation has not been to limit monopoly profits, but, much more dubiously, to divert a portion of those profits to the public purse. We must inquire whether the present character of municipal regulation of cable television is accidental and easily remediable, or the result of more or less inherent conditions of local government. I suspect the latter is the case. First, municipal governments are desperate to find additional sources of revenue. Federal and state income taxation leaves little room for cities to levy their own income taxes. People who enjoy the benefits of municipal services can often escape the cost of those services by moving to the suburbs. And cities act as magnets to poor people who receive, in services and transfer payments, more than they pay in taxes. Since these factors are likely to persist, cities will continue to find cable television an attractive source of additional revenue. Among other factors, an excise tax on cable television may be more progressive than many city taxes (although it could become highly regressive if cable television ever became peculiarly responsive to the needs of the urban poor), and a tax on a new service is likely to be less disruptive of expectations, and therefore less costly politically, than a tax on an existing service.[48] The administrative costs of a tax on cable television are also low, since the municipality collects it from one or at most a few concerns, unlike a property tax or general sales tax.

Second, the traditions of monopoly regulation are very weak at the municipal level, perhaps because the conditions of local government do not favor effective monopoly regulation. Local government does not exhibit the kind of differentiation found at the state and federal levels, where there are regulatory agencies that enjoy a high degree of autonomy and specialization. Rarely will a municipal agency be sufficiently independent to subordinate municipal revenue needs to other welfare dimensions. There is also a problem of scale. It will generally not pay a city to inform itself in adequate detail of the promise and problems of cable television; the question is too large in relation to the impact of the industry on the single city in question. And there is finally a problem of freeloading. The rate at which the cable industry grows nationwide may affect crucially the advent of new services (such as pay televison on the cable) that could have large positive welfare effects; for many of those services, to be remunerative, require the aggregation of audience across many markets. An individual municipality knows, however, that the rate at which cable television spreads in its community will have little effect on the national scene, and that if it retards the local cable company's growth in order to squeeze badly-needed municipal revenues from cable operations, the rate of overall national growth, and the benefits such growth will bring, will not be materially affected.

---

[48] If a new tax is placed on barbers, they feel it in reduced demand for their services, and customers feel it in a higher price for haircuts. A tax on cable television is felt by the cable entrepreneur as an additional cost, but he is rarely a local resident. The tax enters into the fee paid by subscribers; but lacking any experience with a fee that does not include such a tax—having no basis for comparison—they are unlikely to grumble.

Although a continuation of municipal regulation of cable television in its present form will act as a brake on the growth of cable television, it does not follow that municipal regulation should be preempted. Municipalities do need additional sources of revenue and it is not clear that on balance there are many better objects of taxation than cable television. The dilemma is particularly acute for those who believe both that urban problems are serious and that the most exciting prospects for cable television relate precisely to those problems.

The thrust of regulation at the state level has been quite different from that at the municipal level. Several states have passed statutes classifying cable systems as public utilities and regulating entry, rates, and service accordingly.[49] There has been no attempt, as at the municipal level, to derive revenue for the state government from cable television.[50] On a superficial analysis the thrust of state regulation is in the right direction, if what is desired is to encourage the growth of cable television, since the formal objective of public utility regulation is to bring about results akin to those of competition in industries not naturally competitive. Unfortunately, as discussed earlier, there is now strong reason to suspect that public utility regulation works much less well in practice than in theory, and skepticism in this regard is particularly pronounced with respect to traditionally undermanned and underfinanced state public utility commissions.

The thrust of regulation of cable television at the federal level is again different. The federal government, acting through the Federal Communications Commission, has not attempted to regulate cable television rates. The Communications Act of 1934 can perhaps be read as giving the Commission the requisite authority,[51] but it is unlikely to assume it. There are thousands of cable television companies and it would be a formidable undertaking for the Commission to regulate each one of them. It must be aware of the failure of the Federal Power Commission's attempt to regulate the thousands of independent producers of natural gas[52] and cannot desire to repeat that unhappy experience.

The Commission has manifested a dual purpose in its regulation of cable television: to shield over-the-air broadcasting from competitive injury by cable television, and to channel the development of cable television in accordance with the Commission's vaguely perceived, unarticulated, and nonetheless tenaciously maintained vision or master plan for the television industry. The second of these purposes is less well-known than the first. The Commission has long been sensitive, and rightly so, to criticisms that its management of the television industry via control of allocation of the radio spectrum has failed to develop, if indeed it has not stifled, the promise of television.[53] The allocation of spectrum to television adopted in the early 1950s gave most viewers a choice of only a few signals. They would have had a larger choice had the Commission not been wedded to a local-station concept (borrowed from its regulation of the radio industry) whereby each community, other than the smallest, was granted at least one "resident" television station. The price of having local stations was that viewers received fewer signals than if the Commission had permitted large regional stations.[54] The price has proved to be a high one; local stations do very little local programming worthy of the name.[55] The Commission expected the problem to be solved by its allocation of large sections of the UHF band for additional television stations, commercial and noncommercial, but this solution failed—most basically, I believe, because advertising, and *a fortiori* charity, cannot support so many television stations. The costs of over-the-air pay television and a general lack of interest in its potentiality seemed to close off that path to a solution also.

---

[49] The statutes are summarized in New York State Public Service Commission Report [19].

[50] *Ibid.*, p. 112. Vermont has imposed a gross-receipts tax, but only to defray the (presumably small) expenses of the regulatory agency.

[51] See *Columbia Law Review* Note [5].

[52] See MacAvoy [15].

Against this background, cable television appears to the Commission as both danger and possible salvation. It is a danger because it jeopardizes over-the-air broadcasting, in particular the UHF service that the Commission has so assiduously nurtured (though recent scholarship may compel a reevaluation of the threat to UHF).[56] At the same time, cable television offers possible deliverance from the conditions—channel scarcity and dependence on advertiser support—that have seemed to prevent television from fulfilling its promise. The Commission is not unaware of this. But after years of master-planning the evolution of the television industry, it cannot let well enough alone. It has decided to cast free over-the-air television in the role of a medium of popular entertainment that is to be available without charge to all (including those too poor to pay for television and those who live in areas that cannot be wired at reasonable cost), and that is to be supported—necessarily if it is to be a free service—by advertisers. It has decided to cast cable television in the role of innovator. Thus, it has required the larger cable operators to originate programming, hoping that their product will be different from that presented on over-the-air television; has conspicuously permitted them to charge by the program or by the channel for original programming; and has required them to serve as an authentic voice of the local community. The distant-signal limitations that the Commission has imposed on the cable-television industry are also designed, in part, to push the industry away from conventional television and toward presenting more original programs. The Commission is at present mulling over additional rules similarly designed to induce cable television to play a distinctive role in the overall television scene.[57]

---

[53] A polite but penetrating criticism is Johnson [11], pp. 2–8.

[54] The FCC's policy of encouraging a large number of stations creates lots of dead spaces in which channels cannot be utilized because to do so would cause interference with stations broadcasting on the same channel in adjacent areas. That is why many areas of the country can receive only two or three VHF television channels even though 13 such channels have been allocated to VHF television broadcasting.

[55] As indicated by a recent industry-sponsored study. See Land [13].

[56] See Park [22].

Thus far, however, the purpose of protecting over-the-air television from the competition of cable television has been more in evidence than the purpose of shaping cable television into a distinctive service. The limitations that the Commission has placed on the carriage of distant signals have probably retarded the development of cable television by years. Moreover, the capacity of the Commission to shape cable television in affirmative ways is extremely limited. As we have seen, to require cable companies to provide services that they cannot market remuneratively is a dubious method of promoting cable television. But this is the only method of promotion available to the Commission.

We are left with the following alternatives, which unhappily are not mutually exclusive: taxation of cable television by regulation at the municipal level; public utility controls at the state level; and a master-planning approach at the federal level that seeks to carve out a distinctive role for cable television, at the same time seeking to cushion its competitive impact on over-the-air broadcasting. None of the regulatory alternatives is terribly attractive in itself; aggregated, they become distinctly unattractive. The sheer cost of complying with three independent, overlapping, and extensive regulatory schemes could well slow the growth of cable television. And nowhere in the array of regulatory systems do we find any capability for coordinating and evaluating the various regulatory programs.

There is much to be said in these circumstances for a new federal statute defining the respective regulatory roles of federal, state, and local governments. An outline of such a statute will serve to round out our discussion of regulation and to recapitulate earlier proposals of this article.

(1) The statute would place a ceiling on the exactions that municipalities may impose on cable television systems in the form of gross-receipts taxes or free municipal channels, or in any other form, as a condition of granting a franchise. The purpose of this provision would be to accommodate the legitimate needs of the cities for additional sources of revenue with the goal of fostering the development of cable television. Municipal authorities would be permitted—and encouraged—to auction off franchises to the firm promising to provide cable service at the lowest rates to subscribers, as suggested earlier.

(2) States would be permitted, but not required, to regulate cable television companies as common carriers or public utilities. Realism would appear to dictate such a provision, whatever its wisdom, and any bad effects would be mitigated by provision (5)(a) below.

(3) The statute would embody a once-for-all legislative solution to the issues of (a) cable operators' copyright liability and (b) the

---

[57] The Commission's latest thinking is contained in a letter from Dean Burch, Chairman of the FCC, to the Chairmen of the Senate Communications Subcommittee and the House Communications and Power Subcommittee. See [34].

need (if any) to protect over-the-air broadcasting from extinction at the hands of cable television. The FCC seems incapable of resolving either issue; both will have to be resolved if cable television is to be relieved of the onerous restrictions that the Commission has placed upon it. I do not undertake here to suggest the terms of the solutions. An important object here is to facilitate the exit of the FCC from the business of regulating cable television. To this end, Congress might further wish to give permanent statutory embodiment to local-signal carriage and nonduplication rules that the Commission has promulgated and that are no longer within the arena of debate.

(4) The FCC would be forbidden to regulate cable television, with three exceptions:
    (a) The Commission would retain such authority as might be necessary to prevent purely physical interference with FCC licensees. Thus, it could regulate the frequencies used by microwave common carriers that furnish distant signals to cable companies.
    (b) It would retain authority to punish cable operators who violated the police-type regulations that have been imposed on television, such as the rule forbidding the broadcasting of obscene matter.
    (c) It would retain authority to promulgate and enforce reasonable rules limiting cross-ownership of media in the same local market.

(5) An Office of Cable Communications would be established in the federal government. It would have several functions (none regulatory):
    (a) To evaluate state and local regulation of cable television, to advise state and local governments on the uses and modes of regulating cable television, and to submit to Congress proposals for modifying state or local regulatory authority where such a course seemed indicated;
    (b) To serve as a national clearinghouse for information concerning the uses, prospects, regulation, and promotion of cable television; and
    (c) To monitor the implementation of the statute and report to Congress periodically on the progress of cable television and cable communications generally.

The statute would shift the emphasis of governmental intervention in cable communications from regulation to research and evaluation. Such a change in the focus of government activity seems more likely to foster the healthy growth of the industry than a continuation of the present emphasis on proliferating restrictions at many levels of government.

References

1. Alchian, A. A. and Kessel, R. A. "Competition, Monopoly, and the Pursuit of Pecuniary Gain," in *Aspects of Labor Economics*, a Conference Series Report by the National Bureau of Economic Research, Princeton, N. J.: Princeton Univ. Press, 1962, pp. 157–75.

2. Barnett, S. R. "Cable Television and Media Concentration, Part I: Control of Cable Systems by Local Broadcasters." *Stanford Law Review*, Vol. 22, No. 2 (January 1970), pp. 221–329.
3. Coase, R. H. "The Economics of Uniform Pricing Systems." *Manchester School of Economics and Social Studies*, Vol. 15, No. 4 (May 1947), pp. 139–56.
4. ———. "The Problem of Social Cost." *The Journal of Law and Economics*, Vol. 3, No. 2 (October 1960), pp. 1–44.
5. Columbia Law Review "Note: Regulation of Community Antenna Television." *Columbia Law Review*, Vol. 70, No. 5 (May 1970), pp. 837–75.
6. Comanor, W. S. and Mitchell, B. M. "Cable Television and the Impact of Regulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 154–212.
7. Demsetz, H. "Why Regulate Utilities?" *The Journal of Law and Economics*, Vol. 11, No. 1 (April 1968), pp. 55–65.
8. Dirlam, J. B. and Kahn, A. E. "The Merits of Reserving the Cost-Savings from Domestic Communications Satellites for Support of Educational Television." *Yale Law Journal*, Vol. 77, No. 3 (January 1968), pp. 494–519.
9. Gerwig, R. W. "Natural Gas Production: A Study of Costs of Regulation." *The Journal of Law and Economics*, Vol. 5, No. 2 (October 1962), pp. 69–92.
10. Green, M., ed. *The Closed Enterprise System—The Nader Study Group Report on Antitrust Enforcement.* 2 Vols. Washington, D. C.: Center for Study of Responsive Law, 1971.
11. Johnson, L. L. "The Future of Cable Television: Some Problems of Federal Regulation." Rand Memorandum RM-61-99FF, January 1970.
12. Kitch, E. W. "Regulation of the Field Market for Natural Gas by the Federal Power Commission." *The Journal of Law and Economics*, Vol. 11, No. 2 (October 1968), pp. 243–80.
13. Land, Herman W. Associates. *Television and the Wired City.* New York, 1968.
14. Lees, F. A. and Yang, C. Y. "The Redistributional Effect of Television Advertising." *Economic Journal*, Vol. 76, No. 298 (June 1966), pp. 328–36.
15. MacAvoy, P. "The Effectiveness of the Federal Power Commission." *The Bell Journal of Economics and Management Science*. Vol. 1, No. 2 (Autumn 1970), pp. 271–303.
16. McGowan, J., Noll, R., and Peck, M. "Prospects and Policies for CATV." Mimeographed. Washington, D. C.: The Brookings Institution, March 1971.
17. Moore, T. G. "The Effectiveness of Regulation of Electric Utility Prices." *Southern Economic Journal*, Vol. 36, No. 4 (April 1970), pp. 365–75.
18. New York Mayor's Advisory Committee on CATV and Telecommunications. "A Report on Cable Television and Cable Telecommunications in New York City." Mimeographed. New York, September 14, 1968.
19. New York State Public Service Commission. "Regulation of Cable Television by the State of New York." Report to the Commission by Commissioner William K. Jones. Mimeographed. New York, 1970.
20. Ohls, J. C. "Marginal Cost Pricing, Investment Theory and CATV." *The Journal of Law and Economics*, Vol. 13, No. 2 (October 1970), pp. 439–60.
21. Park, R. E. "Cable Television and UHF Broadcasting." Rand Report R-689-MF, January 1971.
22. ———. "Potential Impact of Cable Growth on Television Broadcasting." Rand Report R-587-FF, October 1970.
23. Posner, R. A. "Cable Television: The Problem of Local Monopoly." Rand Memorandum RM-6309-FF, May 1970.
24. ———. "Natural Monopoly and Its Regulation." *Stanford Law Review*, Vol. 21, No. 3 (February 1969), pp. 548–643.
25. ———. "Natural Monopoly and Its Regulation: A Reply." *Stanford Law Review*, Vol. 22, No. 3 (February 1970), pp. 540–46.
26. ———. "Taxation by Regulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 22–50.
27. Robinson, J. *The Economics of Imperfect Competition.* New York: The MacMillan Co., 1933.
28. Stigler, G. *The Organization of Industry.* Homewood, Ill.: Richard D. Irwin, Inc., 1968.

29. *U. S. Code.* Communications Act of 1934, as amended. 47 U.S.C. 307 (1964).
30. U. S. Congress, House, Committee on the Judiciary. *Hearings on H.R. 4347 Before Subcommittee No. 3.* 89th Cong., 1st Session, 1960.
31. U. S. Federal Communications Commission. "Commission Policy on Programming." 25 F.R. 7291 (1960).
32. ———. *First Report and Order in Docket 18397 (CATV).* 20 F.C.C. 2d 201 (1969).
33. ———. *Fourth Report and Order (Subscription Television).* 15 F.C.C. 2d 446 (1968).
34. ———. *In Re Commission Proposals for Regulation of Cable Television,* 31 F.C.C. 2d 115 (1971).
35. ———. *Memorandum Opinion and Order in Docket 18397 (CATV).* 23 F.C.C. 2d 825, 828 (1920).
36. ———. *Notice of Proposed Rulemaking and Notice of Inquiry in Docket No. 18397 (CATV).* 15 F.C.C. 2d 417, 426–27 (1968).
37. ———. *Public Service Responsibility of Broadcast Licensees.* Washington, D. C.: U. S. Govt. Printing Office, 1947.
38. ———. *Second Report and Order in Docket 18397 (CATV).* 23 F.C.C. 2d 816 (1970).
39. ———. *Subscription Television Service.* 3 F.C.C. 2d 1, 21 (1962).

Prospects for cable in the 100 largest television markets

Rolla Edward Park

Economics Department
The Rand Corporation

The Federal Communications Commission, hoping "to get cable moving without jeopardizing over-the-air broadcasting," recently proposed rules that would allow cable systems in the 100 largest television markets to carry a limited number of distant signals. This paper employs a nonlinear model of cable penetration to assess the effect of the proposed rules. The model is developed using a sample of 63 cable systems located where several signals can be received over the air with no particular reception problems. Applying the model to some typical top-100 market situations, the author concludes that the proposed rules "will meet the Commission's objective. But more is necessary to keep cable moving, because the rules by themselves are probably not sufficient to make cable profitable in most of the top-100 markets. To succeed in the cities, cable must attract customers with new services in addition to the traditional package of better reception plus distant signals."

1. Introduction

■ After years of little cable growth in the 100 largest television markets, the Federal Communications Commission has recently outlined rules that would permit cable television systems in these markets to import a limited number of distant signals.[1] Until now, FCC rules have effectively prohibited the importation of any outside signals, while cable operators have maintained that they need outside signals to attract enough customers to operate profitably in most large markets. With the new rules, the Commission hopes "to get cable moving so that the public may receive its benefits and to do so without, at the same time, jeopardizing the basic structure of over-the-air television."[2]

Dr. Park received the B. S. degree in engineering from California Institute of Technology (1957), the M.B.A. from U.C.L.A. (1963), and the Ph.D. in economics from Princeton (1970). He has been with the Economics Department of Rand since 1966, working mainly on airport congestion problems and federal regulation of cable television.

This paper is an abridged version of an earlier report prepared for The Rand Corporation [8]. The work was done under a grant from the John and Mary R. Markle Foundation. The author appreciates the help of Richard Cooper, Leland Johnson, Harriet Porch, and others at The Rand Corporation who commented on earlier versions. The views expressed are those of the author and should not be interpreted as those of the Markle Foundation or The Rand Corporation.

[1] Letter from Dean Burch, Chairman, to the Chairmen of the Senate Communications Subcommittee and the House Communications and Power Subcommittee [13], hereafter cited as "FCC Letter." The markets affected are listed in Appendix A of that letter. The list is derived largely from the American Research Bureau's 1970 prime time households ranking. The rules are to apply within the zone of a 35-mile radius surrounding a specified reference point in each designated community in the market. *Ibid.*, p. 7.

[2] *Ibid.*, p. 5.

The new rules give cable operators at least some of what they have been asking for—some distant signals, if not as many as they would like to have. This is undeniably a boost for cable, but how much of a boost? How many families are likely to subscribe to cable systems operating under the new rules?

These are hard questions. Past experience is at best a tricky guide, because major markets differ in important ways from the places where most cable systems now operate. So far, almost all cable growth has been in places where over-the-air television service is extremely limited or where reception is poor for one reason or another. In contrast, the top-100 markets already have fairly good television service: nearly all have three network stations, and many have other stations as well; and reception of local stations within the 35-mile zones is generally good. Thus, there is usually less reason to subscribe to cable in large markets than in the areas where existing cable systems operate. Therefore, estimates based on the penetration achieved by existing systems are apt to overstate the penetration to be expected in major markets.

Not even those few instances of substantial cable growth in large cities—New York, San Diego, Los Angeles, San Francisco—provide much indication of what to expect in the great majority of the top-100 markets. All of these cities have areas with special reception problems—tall buildings in New York, canyons in San Diego and Los Angeles, hills in San Francisco—and to a large extent the major selling point of cable systems there is improved reception. There are, though, a few dozen cable systems whose experience is indicative of the prospects for cable in most of the top-100 markets. These are systems located where several television signals can be received over the air with no particular reception problem. By carefully choosing and analyzing these systems, I have been able to estimate the fraction of households likely to subscribe to cable in various top-100 market situations once distant signals are imported.

2. Method

■ **Relation to previous studies.** Three previous studies use regression analysis to explain the fraction of households that subscribe to existing cable systems.[3] (I shall frequently refer to the fraction of households that subscribe to cable service as "cable penetration.") But existing systems are, for the most part, located where over-the-air television service is rather poor because of lack of stations or bad reception or both. For this and other reasons the results of these studies are not directly applicable to the typical top-100 markets, where over-the-air service is quite good. The present study uses the same statistical technique but modifies the application in three major ways so that the results *are* relevant for the top-100 markets.

First, the cable systems to be analyzed were carefully chosen to include only those located where three or more (and on the average, over six) television signals can be received over the air with no particular reception problems. A total of 86 systems that seemed to meet these specifications were selected by studying maps. Operators of these systems were then contacted by telephone and questioned to make sure that over-the-air reception is in fact unimpaired in their service areas (except possibly by distance from the transmitters). Sixty-three systems remained in the sample after the telephone screening and are the basis for the analysis reported here.

---

[3] Comanor and Mitchell [1], especially Section 2, "Demand," pp. 175–66; McGowan, Noll, and Peck [4], Appendix A, "The Determinants of CATV Penetration;" and Park [6].

Second, a serious effort was made to ensure that the information on these systems is as accurate as possible. (The three previous studies used published data of questionable accuracy.) The information used in this study is the result of careful questioning of system operators during our telephone conversations with them.

Third, this study takes account of factors that were heretofore neglected, some of which turn out to be very important. For example, the previous studies treat alike all stations that are received over the air. In fact there are important differences in reception quality among different stations, depending on how far away their transmitters are and whether they broadcast on very high frequency (VHF) or ultra high frequency (UHF) channels. Since improved reception is an important reason for subscribing to cable service, these differences have large effects on cable penetration—effects that are taken into account in this study.

☐ **The model.** Denote by S the number of subscribers that a cable system has at some particular time (say, July 1971) and by H the number of homes passed by the system's trunk and feeder cable. By definition, the subscriber penetration achieved by that time, S/H, is equal to some fraction of the penetration it will ultimately achieve:

$$S/H \equiv \text{(ultimate penetration) (maturity factor)}, \qquad (1)$$

where the maturity factor is the fraction of its ultimate penetration that the system has already achieved. I discuss in turn the specification of the two terms on the right-hand side of equation (1).

Ultimate penetration

One expects a cable system's ultimate penetration to depend both on characteristics of the system itself, such as the television signals it carries and the subscription fee it charges, and on characteristics of the environment within which it operates, such as the television signals that can be received over the air and average household income in its franchise area.

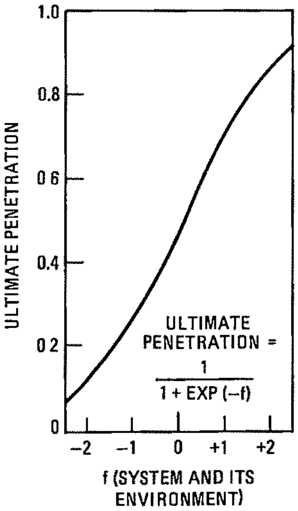
Penetration cannot, of course, be less than zero or more than one; it makes sense to specify a functional form that satisfies these constraints. One reasonable choice is the logit function:

$$\text{ultimate penetration} = \frac{1}{1 + \exp[\,-f(\text{system and its environment})]}. \qquad (2)$$

Over the range of penetration usually observed, expression (2) is roughly linear in f, as shown in Figure 1. For very large values of f, it approaches but cannot exceed one; for very small values, it falls toward zero but never becomes negative.

FIGURE 1

GRAPH OF EQUATION (2)



What exactly are the characteristics of the system and its environment that enter the expression for f? Certainly one important set of characteristics is comparative service on the cable and over the air. With other things held constant, one expects ultimate penetration to be higher for better television service on the cable, and higher for worse television service available over the air for free. Five separate classifications of television service are allowed for:

N, primary network stations:	The nearest stations affiliated with each of the three major commercial networks. (The number of primary network stations cannot exceed three.)
D, duplicate network stations:	Stations, other than primary network stations, affiliated with any of the three major networks.
I, independent stations:	Commercial stations that are not affiliated with any network.
E, educational stations:	Noncommercial stations.
F, Canadian stations:	Stations affiliated with either of the Canadian networks.

I denote the comparative service variables for these five classifications by X_N, X_D, X_I, X_E, and X_F, respectively, and discuss the specification of the *X*s in the following subsection.

Three other potentially important variables are included in the specification of f. One is the price charged by the cable system for its service, denoted by P; one expects lower ultimate penetration for larger P. Another is average household income in the system's franchise area, which I denote by Y; one expects higher ultimate penetration for larger Y. The third is the proportion of households in the system's franchise area that have color television sets, denoted by C; because the improved reception offered by cable is particularly important for color, one expects higher cable penetration for larger C.

Theory does not offer much guidance as to the exact form in which these variables enter f. It is reasonable to suppose that equal proportionate changes (rather than equal absolute changes) in the variables have equal effects on ultimate penetration, and this suggests that they enter logarithmically.[4] Also, considerations of ease of estimation dictate a linear form. Consequently, I specify

$$\begin{aligned} f(\text{system and its environment}) = \beta_0 + \beta_1 \log(X_N) + \beta_2 \log(X_D) \\ + \beta_3 \log(X_I) + \beta_4 \log(X_E) + \beta_5 \log(X_F) + \beta_6 \log(P) \quad (3) \\ + \beta_7 \log(Y) + \beta_8 \log(C) + \epsilon, \end{aligned}$$

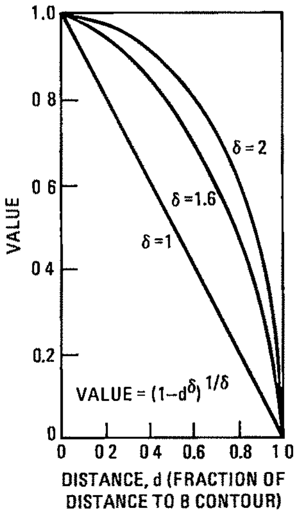
where ϵ is a random error term that represents the influence on ultimate penetration of factors that I am unable to include in the model.

---

[4] I realized after completing the calculations that it would be better to enter C linearly. Using log (C) implies that S/H goes to zero as C goes to zero, which is certainly not true. Because C in my sample is never less than 0.37, the mistake is not serious enough to warrant redoing the calculations.

FIGURE 2

DISTANCE DISCOUNT FUNCTION



Comparative service variables, X_i

For use as comparative service variables, one needs some sort of ratio of service on the cable to service over the air. One form used in previous studies is $X_i = (1 + W_i)/(1 + B_i)$, where W_i is the number of stations of type i carried on the cable and B_i is the number of stations of type i receivable over the air. The weakness of this formulation is that it does not take into account the fact that some over-the-air signals are better than others. Close to the station's transmitter, pictures received over the air may be just as good as those on the cable. Out near the B contour,[5] they may be much worse. Yet this formulation counts them the same.

In counting up over-the-air service in the model used here, I discount the value of stations depending on the distance of their transmitters from the cable system in question. Expressing distance, d, as a fraction of the distance out to the B contour, each station is valued at $(1 - d^\delta)^{1/\delta}$, where δ is a parameter to be estimated. This functional form allows station values to decline initially less rapidly than distance increases, as shown in Figure 2.

Stations are assumed to be valued at zero at and beyond their B contours. Ideally, of course, one would want to estimate the distance at which value goes to zero, rather than assume it. As a practical matter, though, it is not feasible to estimate more than one parameter in a distance discount function.

Also, it seems likely that UHF stations are of less value over the air than are VHF stations, for a variety of reasons: some households have older television sets that cannot receive UHF signals at all; many do not have the special antenna that is necessary in many areas for satisfactory UHF reception; even with a UHF antenna, reception may not be as good as for VHF stations; finally, the UHF tuner is less convenient to use than the VHF tuner.[6] Consequently, a UHF station is valued at the fraction γu as much as a VHF station the same distance away, where γ is a parameter to be estimated and u is the fraction of households in the cable system's service area that have UHF receivers.

Pulling all of this together, I express the comparative service variables X_i as

$$X_i = \frac{1 + W_i}{1 + \gamma u \sum_{U_i} (1 - d^\delta)^{1/\delta} + \sum_{V_i} (1 - d^\delta)^{1/\delta}}, \qquad (4)$$

$$i = N, D, I, E, F,$$

where U_i is the set of B-contour or better UHF stations in class i and V_i is the set of VHF stations.

Maturity factor

Ultimate penetration, as given by equations (2) through (4), is affected by many variables. In contrast, I make the simple hypothesis that the maturity factor in equation (1) is affected only by the age of the system. That is, all cable systems are assumed to follow similar growth paths, with each reaching the same fraction of its ultimate penetration at the same age.

---

[5] B contours are calculated using formulas specified by the FCC. They are intended to enclose areas within which the quality of the picture is satisfactory to the median observer at least 90 percent of the time for at least 50 percent of receiving locations.

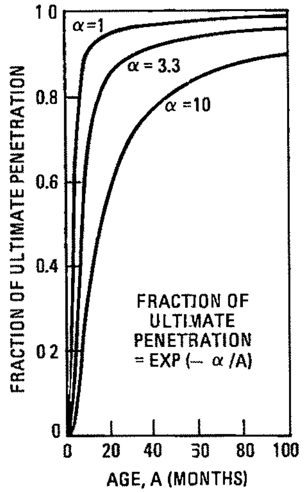
[6] For evidence that UHF stations attract smaller audiences than do VHF stations broadcasting the same programs, even considering only households that have UHF receivers, see Park [5].

As in my previous study,[7] I initially hypothesize that cable systems grow along a logistic path such as those shown in Figure 3. Along a logistic path, a system of age A has achieved the fraction $\exp(-\alpha/A)$ of its ultimate size, where α is a parameter to be estimated, so I specify

$$\text{maturity factor} = \exp(-\alpha/A)\,. \tag{5}$$

FIGURE 3

LOGISTIC GROWTH PATHS



The equation to be estimated

Manipulating equations (1) through (5), we obtain

$$\log\left(\frac{S/H}{\exp(-\alpha/A) - S/H}\right) = \beta_0 + \beta_1\log(X_N) + \beta_2\log(X_D) + \beta_3\log(X_I) + \beta_4\log(X_E) + \beta_5\log(X_F) + \beta_6\log(P) + \beta_7\log(Y) + \beta_8\log(C) + \epsilon\,, \tag{6}$$

where the Xs are given by equation (4). Assuming that the error term ϵ is well behaved, one may estimate the unknown parameters in equations (4) and (6), $\alpha, \gamma, \delta, \beta_0, \beta_1, \cdots \beta_8$, by nonlinear least squares if one has the necessary data for an appropriate sample of cable systems.

☐ The data.

Sample selection

The sample is made up of systems located where several stations are available over the air without any particular reception problems (other than those due to distance from the transmitters). To choose the sample, I used maps in the *CATV Atlas*[8] to select systems that are inside the A contours[9] of three or more television stations and are not in locations with obvious reception problems. In a few cases, systems were included even though they failed the three-A-contour test if they were close to the A contours and were within the B contours of three primary network stations. In all, 86 systems were chosen in this way.

To determine whether these systems are in fact located where over-the-air reception is no problem, a representative of each system was called on the telephone and asked the following questions: How would you describe over-the-air reception in your service area? Is there any reason other than distance that reception is impaired? What do you consider to be the most important reason and the second most important reason that people sign up for your service? If the answers to these questions revealed any particular reception problem (not just distance), that system was eliminated from the sample. Fourteen systems were eliminated because of reception problems, and two others because they could not be reached by telephone. Later on,

---

[7] See [6].

[8] See [9], 1969–1970 edition.

[9] A contours are intended to enclose areas within which picture quality is satisfactory to the median observer at least 90 percent of the time for at least 70 percent of receiving locations.

TABLE 1

OVER–THE–AIR SIGNALS IN SAMPLE SYSTEMS' SERVICE AREAS

	AVERAGE	STANDARD DEVIATION	MINIMUM	MAXIMUM
TOTAL NUMBER OF A–CONTOUR OR BETTER SIGNALS	4.14	1 40	2[a]	8
TOTAL NUMBER OF B–CONTOUR OR BETTER SIGNALS	6.17	2 41	3	12
PRIMARY NETWORK SIGNALS, A–CONTOUR OR BETTER	2.79	0.45	1[a]	3
PRIMARY NETWORK SIGNALS, B–CONTOUR OR BETTER	2 95	0 21	2[b]	3

[a] IN ONE CASE, WHICH HAS THREE PRIMARY NETWORK SIGNALS, B–CONTOUR OR BETTER.

[b] IN THREE CASES, ALL OF WHICH HAVE A TOTAL OF THREE OR MORE A–CONTOUR SIGNALS.

seven more systems were eliminated because of various data problems, leaving 63 systems in the final sample.

Table 1 summarizes over-the-air service in the communities served by these 63 systems. The average system is located inside the *A* contour of slightly more than four stations. All except one are located inside the *B* contours of a complete lineup of network stations (three primary network signals). These levels of service are roughly the same as those found in most of the top-100 markets.

Data sources and quality control

Two types of data are needed to estimate equation (6): data on the cable systems themselves, and data on the environments in which they operate.

Data on the cable systems. Previous studies have all used the *Television Factbook*[10] as their source of data on cable systems. All of the required data are usually, but not always, present in the *Factbook* listings: number of subscribers and number of homes in front of the cable, from which penetration, S/H, is calculated; installation fee and monthly subscriber fee, from which annual price of cable service, P, is calculated;[11] date service began, from which the age of the system in months at the time of the observation, A, is calculated; and a list of television stations carried by the system, from which the numbers of different kinds of signals, W_i, are derived for use in the numerators of the comparative service variables.

In quite a few cases, however, some or all of the necessary data are missing from the *Factbook* listings. In other cases, there are obvious inconsistencies between the data given in successive issues (for example, different begin-service dates are sometimes given). Furthermore, data for a given year are sometimes internally inconsistent (for example, the number of subscribers sometimes exexceeds the number of homes passed by the cable). Consequently, I decided to obtain data for this study directly from system operators.

---

[10] See [10].

[11] As 12 (monthly fee) + 0.2 (installation fee), following the study by McGowan, Noll, and Peck [4], the rationale being that families move every five years on the average so that installation fee should be amortized over that period. Since 0.2 (installation fee) is typically less than 5 percent of the total so calculated, the exact assumption as to how installation fee enters is not too important.

To obtain better quality data, system representatives were questioned during the telephone survey mentioned earlier. Current figures (as of July 1971) were obtained for all of the system data, including a complete list of cable channel usage (what station or stations are carried on each channel). Inconsistencies with the *Factbook* listings were pointed out, discussed, and resolved if possible. Complete and satisfactory current data were obtained for 63 systems, which constitute the final sample.

A problem in measuring penetration arises because of the existence of multiple dwellings such as apartment buildings and hotels. Ideally, one would define penetration as number of occupied dwelling units served divided by number of occupied dwelling units passed. For example, a fully occupied 50-unit apartment house that is completely wired would count as 50 in both the numerator and denominator of the expression for penetration. Unfortunately, cable operators are not always able or willing to estimate these quantities.

The difficulty arises largely because of the practice of signing up multiple dwellings as bulk-rate customers. A whole apartment house served in this way appears on the company's records as a single subscriber. Cable companies with bulk-rate customers frequently have no record (or at least no convenient record) of the number of individual dwelling units they serve. Also, since these operators are unaccustomed to thinking in terms of individual dwelling units, they are often unwilling even to estimate the number of individual units that they serve or that their cable passes.

Short of making detailed field surveys, there is no completely satisfactory way to solve this problem. The makeshift procedure used was first to try to get estimates of individual units passed and served. If that failed, an attempt was made to ensure that the estimates of subscribers and homes passed were consistent, so that, for example, if an apartment house on bulk rate were counted as a single subscriber, it was also counted as a single home passed.

Data on cable systems' environments. Data on the environmental variables necessary to estimate equation (6) come from a variety of sources. Television stations receivable over the air in the community served by the cable system are taken from maps in the *CATV Atlas*[12] and include all stations that put a B contour over the community. Distance from the transmitter is scaled from these maps, supplemented by those in the Stations Volume of the *Factbook*, and expressed as a fraction, d, of the distance from the transmitter to the B contour, measured through the community. As proxies for the proportion of families in the system's service area that have UHF receivers, u, and color sets, C, figures were used for the television market within which the system is located.[13] Data on income, Y, are constructed from 1960 census figures for median household income by county for 1959.[14] These are inflated to estimated 1969 levels in proportion to the increase in median family income from 1959 to 1969 for the United States as a whole, as reported in *Current Population Reports.*[15] The year 1969 is the latest for which an official income estimate is available; further inflation to 1971 levels uses an assumed factor of 1.1. This completes the data necessary to estimate equation (6).

---

[12] See [9], 1970–1971 edition.

[13] These figures are reported in the *Factbook* [10], 1971–1972 edition, as of a February/March 1971 Survey.

[14] See *County and City Data Book 1967* [11].

☐ **The estimated equations.** Using the data described above, I estimated the unknown parameters in equation (6) using a least-squares procedure involving nested search on the nonlinear parameters α, γ, and δ.[16] The basic procedure follows:

(1) Set $\delta = 1.5$.
(2) Calculate the over-the-air service summations in equation (4). Set $\alpha = 2$.
(3) Run a series of linear regressions for trial values of γ, converging on the value of γ that minimizes the sum of squared residuals (for given α and δ). Denote that value of γ by $\gamma^*(\alpha,\delta)$, the corresponding estimates of β_i by $\beta_i^*(\alpha,\delta)$, and the minimal sum of squared residuals by $SSR^*(\alpha,\delta)$.
(4) Repeat step 3 for $\alpha = 4$ and $\alpha = 6$.
(5) Fit a quadratic through the three observed α, $SSR^*(\alpha,\delta)$ points. On that curve, find the value of α (to two significant figures) that minimizes $SSR^*(\alpha,\delta)$. Denote that value of α by $\alpha^*(\delta)$ and the corresponding sum of squared residuals by $SSR^*(\delta)$.
(6) Repeat steps 2 through 5 for $\delta = 1.3$ and $\delta = 1.7$.
(7) Fit a quadratic through the three observed δ, $SSR^*(\delta)$ points. On that curve, find the value of δ (to two significant figures) that minimizes $SSR^*(\delta)$. Denote that value of δ by δ^*.
(8) Set $\delta = \delta^*$ and do steps 2 through 5.
(9) Set $\alpha = \alpha^*(\delta^*)$ and do step 3.
(10) The final values for δ^*, α^*, γ^*, and β_i^* are approximate least-squares estimators of the unknown parameters in equations (4) and (6).[17]

The full equation: line (1)

The regression equation estimated as described is shown on line (1) in Table 2. The explanatory power is quite high: R-squared equals 0.734. All except one of the parameter estimates have the right sign and reasonable values.

---

[15] See [12].

[16] This was easier than it might have been because of the availability of a versatile regression analysis program written by William J. Raduchel of Harvard University.

[17] Additional searching in the vicinity of some of the equations estimated this way indicates that they are very good approximations to the true least-squares estimates. The calculated sums of squared residuals seemed generally to be within 1/20 of one percent of the true minimum. Thus the sums of squared residuals reported in Table 2 are generally accurate to the number of significant digits shown. The estimated coefficients may differ slightly from their true least-squares values but generally only in the last digit reported. Individual coefficients are important only in lines (1) and (*) of the table; for these two lines, additional searching confirms that no other estimates with α and δ rounded as shown have lower squared residuals.

TABLE 2

ESTIMATED EQUATIONS[a]

LINE	AGE DISCOUNT PARAMETER (α)	UHF VALUE FRACTION (γ)	DISTANCE DISCOUNT PARAMETER (β)	CONSTANT (β_0)	COMPARATIVE SERVICE					PRICE (β_6)	INCOME (β_7)	COLOR SET OWNERSHIP (β_8)	SQUARED RESIDUALS	R–SQUARED[b]	STANDARD ERROR OF ESTIMATE	DEGREES OF FREEDOM
					PRIMARY NETWORK (β_1)	DUPLICATE NETWORK (β_2)	INDEPENDENT (β_3)	EDUCATIONAL (β_4)	CANADIAN (β_5)							
(1)	3.3	0 726	1.6	−8.291 (−1 99)	3 278 (7.40)	0 252 (2 44)	0.201 (1.57)	0.320 (2.29)	−0.503 (−2.08)	−1.373 (−2 19)	1.377 (4.23)	0 542 (1.11)	7.578	0 7337	0.3855	51
(2)	0.0[c]	0.708	1.5	−9.841 (−2 34)	3.094 (7.45)	0.299 (2.84)	0.233 (1 81)	0 250 (1.77)	−0.481 (−2 01)	−1 201 (−1.91)	1 432 (4.38)	0.332 (0 67)	7.774	0.7269	0.3867	52
(3)	2.3	0.000[c]	0.7	−15.14 (−2.79)	0.858 (4.22)	0.510 (3 34)	0.119 (0.76)	0.202 (1.16)	−0.550 (−2.24)	−0.531 (−0.68)	1.671 (4.27)	0 189 (0.29)	11.360	0 6009	0.4674	52
(4)	3.3	1 000[c]	1 4	−6.398 (−1 39)	2 373 (5.58)	0.407 (3 53)	0.282 (2.08)	0.173 (1.12)	−0.654 (−2.49)	−1.469 (−2.11)	1.255 (3.47)	0 962 (1 82)	9 528	0 6652	0.4281	52
(5)	6.0	0.720	∞[c]	1.204 (0.22)	2 607 (3.15)	−0.081 (−0.64)	0.090 (0.53)	0.638 (3.50)	−0.245 (−0.67)	−1 582 (−1 71)	0.646 (1.50)	2 001 (3 02)	15.232	0 4648	0 5412	52
(6)	2 5	0.572	1.0[c]	−13 419 (−2.93)	1.674 (6 02)	0 474 (3.99)	0.236 (1.71)	0.199 (1 29)	−0 569 (−2.48)	−0 813 (−1.23)	(1.616) (4.69)	0.200 (0 38)	8 554	0 6995	0 4056	52
(7)	4.0	0.756	1 7	0.000[c]	3.317 (7 05)	0 197 (1.92)	0 224 (1.74)	0.389 (2.80)	−0.448 (−1.82)	−2184 (−4.46)	0.901 (4 06)	1 171 (2 87)	8 094	0 7156	0 3945	52
(8)	3.5	4 355	0 5	−9 619 (−1.77)	0 000[c]	0.611 (4 14)	0.383 (2.98)	0 157 (1 44)	−0.520 (−1.91)	−0.474 (−0 56)	1.180 (3 00)	1 208 (1.94)	13 452	0.5263	0 5086	52
(9)	4.1	0.706	1.7	−6.117 (−1.45)	3.576 (7.61)	0.000	0 169 (1 28)	0 306 (2 75)	−0.286 (−1.23)	−1.240 (−1 92)	1.117 (3.54)	0 652 (1.29)	8.293	0 7086	0 3994	52
(10)	3 9	0.715	1 6	−9.086 (−2.17)	3.302 (7.43)	0 234 (2 23)	0.000[c]	0 351 (2.48)	−0 552 (−2.27)	−1 496 (−2.37)	1 547 (4.99)	0.747 (1.56)	7.930	0.7214	0 3905	52
(11)	1.8	0 721	1.4	−11.583 (−2.26)	2.972 (7.70)	0.336 (3.10)	0.239 (1.82)	0.000[c]	−0.549 (−2.27)	−1.120 (−1.74)	1 595 (4.90)	0 337 (0 67)	8.143	0.7139	0 3957	52
(12)	2.7	0.705	1.6	−7.311 (−1.72)	3.310 (7.35)	0 164 (1.67)	0.237 (1.81)	0.336 (2.33)	0 000[c]	−1.106 (−1.74)	1 138 (3.68)	0 401 (0.80)	8 196	0.7120	0.3970	52
(13)	2 3	0.679	1.4	−15.540 (−4.77)	2.850 (7 27)	0 292 (2.68)	0 245 (1.85)	0 249 (1.70)	−0 432 (−1 80)	0.000[c]	1.487 (4.42)	−0 027 (−0 06)	8.207	0.7116	0 3973	52
(14)	3.7	0 777	1 8	5.530 (1.76)	3 310 (5 92)	0 067 (0 63)	0.365 (2.70)	0.470 (3.08)	−0.136 (−0 53)	−1 458 (−2.03)	0 000[c]	1.421 (2.74)	10.051	0.6468	0 4396	52
(15)	2.6	0.694	1 5	−11.356 (−3 31)	3.106 (7 80)	0.282 (2.70)	0.238 (1 93)	0.291 (2 07)	−0.480 (−2.02)	−1.075 (−1 85)	1 521 (4.96)	0.000[c]	7.702	0 7294	0 3849	52
(*)	18.0	0.731	1.6	−8 159 (−2 03)	3.098 (7 21)	0 290 (2 90)	0 212 (1 72)	0 298 (2.20)	−0.540 (−2 31)	−1 473 (−2 42)	1.398 (4.44)	0 523 (1.11)	7 089	0 7509	0 3728	51

[a] VALUES SHOWN IN PARENTHESES ARE t VALUES CONDITIONAL ON THE VALUES LISTED FOR NONLINEAR PARAMETERS BUT CORRECTED FOR DEGREES OF FREEDOM USED IN ESTIMATING THE NONLINEAR PARAMETERS

[b] FRACTION OF THE VARIANCE OF LOG $\left(\frac{S/H}{1-S/H}\right)$, WHICH EQUALS 28.461, EXPLAINED BY THE ESTIMATED EQUATION.

[c] ASSUMED, NOT ESTIMATED

The age discount parameter, α, is estimated to be 3.3, indicating that systems approach maturity quite rapidly, as shown by the middle curve in Figure 3. The UHF value fraction, γ, estimated to be 0.73, indicates that UHF stations do indeed suffer from the sorts of handicaps, in addition to incomplete UHF set penetration, that are discussed above.[18] The estimated value of the distance discount parameter, δ equals 1.6, indicates that reception quality declines with distance in about the expected manner, as shown by the middle curve in Figure 2.

The estimated coefficients of the comparative service variables, β_1 through β_5, indicate that improvement in primary network service is by far the most important reason that people subscribe to cable.

---

[18] Page 134 *infra.*

"Improvement" here does not necessarily mean carrying a network that cannot be received over the air, because almost all of the systems in the sample are located where all three networks *can* be received without cable. Instead, it means improvement in reception of primary network stations that are located some distance away, or that broadcast on UHF channels. The effects of improved duplicate network, independent, and educational services are all positive but much smaller than the effect of primary network service. The estimated coefficient for Canadian service has the wrong sign.

The coefficients for price, β_6, and income, β_7, are almost the same in absolute value, indicating that a one-percent decrease in price of cable service increases penetration by about the same amount as would a one-percent increase in income. These coefficients cannot be directly interpreted as price and income elasticities. Appropriate translation, though, shows that penetration is about unit elastic with respect to both price and income.[19]

The coefficient of color set penetration, β_8, is positive, as expected, but considerably smaller than those of price, income, and primary network service.

Significance tests: lines (2) through (15)

Student's t values for the linear β coefficients, corrected for the extra degrees of freedom used in estimating the nonlinear parameters, are given in parentheses in Table 2. But these provide nothing more than a very rough guide to the relative statistical significance of the β coefficients, because they are conditional on the listed values of the nonlinear parameters.

To test the significance of *all* of the parameters, both linear and nonlinear, I estimated the equations on lines (2) through (15). For line (2), I set the age discount parameter α equal to zero and estimated all of the others using a straightforward modification of the algorithm given above. Setting α equal to zero assumes that age does not matter, that cable systems reach their ultimate penetration instantaneously. The question of whether α is significantly different from zero is equivalent to the question of whether line (1) explains significantly more variance than does line (2). This is tested by the first analysis of variance shown in Table 3. Surprisingly, we find that the additional variance explained by equation (1) could easily be the result of chance if α were zero; α is not significantly different from zero at anything approaching the 0.05 level. I return to this surprising result in the next subsection.

Line (3)[20] in Table 2 and the associated analysis of variance in Table 3 show that the UHF value fraction γ is significantly different from zero at well beyond the 0.01 level. It may be of even more in-

---

[19] Using equations (2) and (3), one finds that income elasticity,

$$\frac{Y}{S/H}\cdot\frac{d(S/H)}{dY} = \beta_7\left(\frac{\exp(-f)}{1+\exp(-f)}\right).$$

The average value of penetration in my sample is about 0.3, and the corresponding value of f from Figure 2 is -0.8. Substituting for f and β_7 in the above expression, income elasticity is calculated to be 0.96. In a similar way, price elasticity is calculated to be 1.01.

[20] For line (3) and line (8), trial values for δ of 0.4, 0.8, and 1.0 were used in the algorithm given above.

TABLE 3

ANALYSES OF VARIANCE

SOURCE	NULL HYPOTHESIS	FRACTION OF VARIANCE	DEGREES OF FREEDOM	MEAN FRACTION	P–STATISTIC
LINE (2)	$\alpha = 0$	0.7269	10		
ADDED BY LINE (1)		0.0068	1	0.0068	1.30
LINE (3)	$\gamma = 0$	0.6009	10		
ADDED BY LINE (1)		0.1328	1	0.1328	25.43[a]
LINE (4)	$\gamma = 1$	0.6652	10		
ADDED BY LINE(1)		0.0685	1	0.0685	13.12[a]
LINE (5)	$\delta = \infty$	0.4648	10		
ADDED BY LINE (1)		0.2689	1	0.2689	51.50[a]
LINE (6)	$\delta = 1$	0.6995	10		
ADDED BY LINE (1)		0.0342	1	0.0342	6.55[b]
LINE (7)	$\beta_0 = 0$	0.7156	10		
ADDED BY LINE (1)		0.0181	1	0.0181	3.47
LINE (8)	$\beta_1 = 0$	0.5263	10		
ADDED BY LINE (1)		0.2074	1	0.2074	39.73[a]
LINE (9)	$\beta_2 = 0$	0.7086	10		
ADDED BY LINE (1)		0.0251	1	0.0251	4.81[b]
LINE (10)	$\beta_3 = 0$	0.7214	10		
ADDED BY LINE (1)		0.0123	1	0.0123	2.36
LINE (11)	$\beta_4 = 0$	0.7139	10		
ADDED BY LINE (1)		0.0198	1	0.0198	3.79
LINE (12)	$\beta_5 = 0$	0.7120	10		
ADDED BY LINE (1)		0.0217	1	0.0217	4.16[b]
LINE (13)	$\beta_6 = 0$	0.7116	10		
ADDED BY LINE (1)		0.0221	1	0.0221	4.23[b]
LINE (14)	$\beta_7 = 0$	0.6468	10		
ADDED BY LINE (1)		0.0869	1	0.0869	16.64[a]
LINE 15	$\beta_8 = 0$	0.7294	10		
ADDED BY LINE (1)		0.0043	1	0.0043	0.82
LINE (1)		0.7337	11		
UNEXPLAINED		0.2663	51	0.00522	
TOTAL		1.0000	62		
LINE (2)	$\alpha = 0$	0.7269	10		
ADDED BY LINE (*)		0.0240	1	0.0240	4.91[b]
LINE (*)		0.7509	11		
UNEXPLAINED		0.2491	51	0.00488	
TOTAL		1.0000	62		

[a] SIGNIFICANT AT THE 0.01 LEVEL, TEST STATISTIC IS F(1, 51, 0.01) = 7.16.

[b] SIGNIFICANT AT THE 0.05 LEVEL, TEST STATISTIC IS F(1, 51, 105) = 4.03

terest to know whether γ is significantly different from one—whether UHF stations are significantly different from VHF stations in their effect on cable penetration, even after adjusting for incomplete UHF set penetration. This is tested in line (4) of Table 2 and the corresponding analysis of variance in Table 3, which show that γ is indeed different from one, again at well beyond the 0.01 significance level.

Line (5) assumes that there is no distance discount. In counting up over-the-air service, a station is valued the same all the way out to its *B* contour, just as in the three previous studies.[21] (In terms of the discount function, no distance discount is equivalent to assuming

---

[21] Cited above, note 3.

δ to be infinitely large.) Line (5) performs very poorly compared to line (1). Discounting stations for distance increases R-squared by over 25 percentage points and, as shown by the corresponding analysis of variance in Table 3, is significant way beyond the 0.01 level. This confirms that the previous studies were deficient in not taking distance into account.

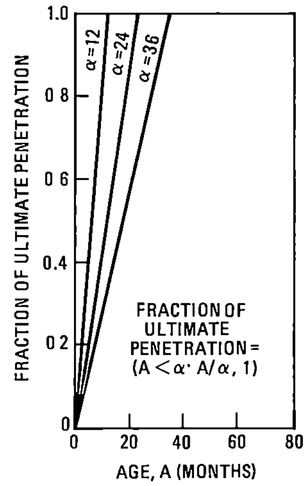
On the other hand, a curved distance discount function (δ equals 1.6) is somewhat less strongly compelled by the data in preference to, say, a linear one (δ equals one). This is shown by line (6) in Table 2 and the corresponding analysis of variance in Table 3. Estimating δ rather than assuming it equal to one explains about 3½ more percentage points of variance and barely misses being significant at the 0.01 level (is solidly significant at the 0.05 level).

Lines (7) through (15) and the associated analyses of variance test the significance of the β parameters. Coefficients of comparative primary network service and income are significant at the 0.01 level. (Primary network service adds nearly as much to the explanatory power of the equation as does discounting for distance.) Coefficients of duplicate network service, Canadian service, and subscription price are all significant at the 0.05 level (but Canadian service has the wrong sign). The other coefficients are not significant at the 0.05 level.

A revised maturity factor: line (*)

FIGURE 4

STRAIGHT LINE GROWTH PATHS



There is a very strong *a priori* presumption that older, more mature cable systems have higher penetrations than young ones. So the finding above that the age discount parameter α estimated in equation (6) is not significantly different from zero is disconcerting. One reason for this surprising result seems to be that the actual growth path is so steep, and the number of very young systems in my sample so small, that distinguishing between the actual growth path and an instantaneous one is difficult. Another reason may be that the maturity factor is not correctly specified in equation (5).

Consequently, I respecify a straight line growth path (Figure 4):[22]

$$\text{maturity factor} = (A < \alpha\colon\ A/\alpha;\ 1) \qquad (5')$$

and estimate the resulting equation:

$$\log\left(\frac{S/H}{(A < \alpha\colon A/\alpha;\ 1) - S/H}\right) = \beta_0 + \beta_1 \log(X_N) + \beta_2 \log(X_D) + \beta_3 \log(X_I) + \beta_4 \log(X_E) + \beta_5 \log(X_F) + \beta_6 \log(P) + \beta_7 \log(Y) + \beta_8 \log(C) + \epsilon, \qquad (6')$$

where the Xs are given by equation (4).

The estimated coefficients[23] are shown as line (*) in Table 2. Cable systems are estimated to attain maturity at 18 months (α equals 18). The other estimated coefficients are little changed from line (1). Line (*) explains almost 2 percentage points more variance than does line (1), and almost 2½ more than line (2), which assumes no age discount. This additional variance is significant at the 0.05 level, as shown by the last analysis of variance in Table 3; in line (*), α *is* significantly different than zero.

---

[22] Meaning that if A is less than α, maturity factor equals A/α, and otherwise, maturity factor equals 1.

[23] Obtained using trial values of α = 16, 18, and 20 in the algorithm given above.

Line (*) estimates the preferred equation that I use for calculating expected penetration in the next section:

$$\log\left(\frac{S/H}{(A < 18^b\colon A/18^b;\ 1) - S/H}\right) = -\ 8.159 + 3.098^a \log(X_A)$$
$$+\ 0.290^b \log(X_D) + 0.212 \log(X_I) + 0.298 \log(X_E) \qquad (*)$$
$$-\ 0.540^b \log(X_F) -\ 1.473^b \log(P) + 1.398^a \log(Y)$$
$$+\ 0.523 \log(C)\,,$$

where

$$X_i = \frac{1 + W_i}{1 + 0.731^a u \sum_{U_i} (1 - d^{1.6a})^{1/1.6a} + \sum_{V_i} (1 - d^{1.6a})^{1/1\ 6a}}\,,$$
$$i = N, D, I, E, F.$$

The superscripts of the coefficients in equation (*) indicate significance levels (*a* for the 0.01 level, *b* for the 0.05 level), on the assumption that coefficients found to be significant in line (1) remain significant in line (*). Comparison of the conditional *t* values for the two lines suggests that this assumption is conservative.

3. Results

■ Using the relationships found for the cable systems in my sample,[24] this section presents estimates of cable penetration to be expected under the FCC's proposed rules in some typical top-100 market situations. Generally, expected penetration at the center of the market ranges from about 20 to 35 percent; at the edges of the 35-mile zone, it ranges from about 30 to 60 percent.

☐ **The FCC's proposed rules.** The rules outlined in the FCC's letter[25] specify stations that cable systems *must* carry and others that they *may* carry. Very briefly summarized, the rules are as follows:

(1) A cable system *must* carry all stations licensed to communities within 35 miles of the system.
(2) A cable system *must* carry stations that are significantly viewed in the system's community.
(3) A cable system *may* carry other commerical stations as necessary to provide these minimum service levels:
    (a) In the top 50 markets, three network stations and three independents;
    (b) In the second 50 markets, three network stations and two independents.
(4) A cable system *may* carry two commercial stations in addition to those it must carry. Stations carried to provide minimum service count against this allowance.
(5) A cable system *may* carry an unlimited number of noncommercial (educational) stations and foreign-language stations in the absence of objection by interested parties.

---

[24] Specifically, equation (*) *infra*.

[25] FCC Letter [13], pp. 5–25.

TABLE 4

EXPECTED CABLE PENETRATION IN TOP–100 MARKETS: SOME HYPOTHETICAL EXAMPLES[a]

LINE'	LOCAL SIGNALS[b]	OVERLAPPING SIGNALS[b] ON CABLE	PENETRATION[c] (PERCENT) WITHOUT DISTANT SIGNALS		DISTANT SIGNALS[b] ON CABLE	PENETRATION[c] (PERCENT) WITH DISTANT SIGNALS	
			MIDDLE[d]	EDGE[d]		MIDDLE[d]	EDGE[d]
(1)	3NV, 1IV, 2IU, 1EV	NONE	15 (10–23)	26 (17–36)	2I, 2E	20 (13–28)	32 (22–43)
(2)	3NV, 1EU	NONE	16 (10–23)	25 (17–35)	3I, 1E	22 (15–31)	34 (24–45)
(3)	2NV, 1NU, 1EU	NONE	20 (13–28)	30 (21–41)	2I, 1E	26 (18–36)	38 (27–50)
(4)	1NV, 2NU, 1EU	NONE	23 (16–33)	35 (25–46)	2I, 1E	30 (21–41)	43 (32–55)
(5)	3NU, 1EU	NONE	27 (19–38)	39 (28–51)	2I, 1E	35 (25–46)	48 (36–60)
(6)[e]	3NV, 1EU	NONE	20 (14–29)	31 (22–43)	3I, 1E	28 (19–39)	41 (30–53)
(7)[f]	3NV, 1EU	NONE	30 (21–41)	43 (32–55)	3I, 1E	39 (28–51)	54 (42–65)
(8)	3NV, 1EU	3NV	18 (12–26)	26 (18–37)	3I, 1E	25 (17–35)	35 (25–47)

[a] UHF SET PENETRATION USED IN THE CALCULATIONS DEPENDS ON THE NUMBER OF LOCAL NETWORK UHF SIGNALS:

LOCAL NETWORK UHF SIGNALS	0	1	2	3
UHF SET PENETRATION:	0 80	0.90	0 95	0.99 .

EXCEPT AS NOTED FOR LINES (6) AND (7), ANNUAL PRICE OF CABLE SERVICE IS $63 AND MEDIAN FAMILY INCOME IS $10,000. THROUGHOUT, COLOR SET PENETRATION IS 0 5.

[b] N MEANS NETWORK, I MEANS INDEPENDENT , E MEANS EDUCATIONAL; V MEANS VHF, U MEANS UHF.

[c] THE TOP FIGURE IS THE MOST LIKELY VALUE THE FIGURES IN PARENTHESES SHOW AN APPROXIMATE 80–PERCENT CONFIDENCE INTERVAL; THUS THERE IS ABOUT 1 CHANCE IN 10 THAT PENETRATION WILL FALL SHORT OF THE LOWER FIGURE, AND 1 CHANCE IN 10 THAT IT WILL EXCEED THAT HIGHER FIGURE

[d] SEE THE MAP (FIGURE 5) FOR LOCATIONS

[e] FOR LINE (6) ONLY, MEDIAN FAMILY INCOME IS $12,500

[f] FOR LINE (7) ONLY, ANNUAL PRICE OF CABLE SERVICE IS $36, OR $3 PER MONTH

☐ **Expected cable penetration: hypothetical examples.** Table 4 shows the effects on cable penetration to be expected when the FCC's proposed rules are applied in some typical top-100 market situations. Let us look first at the second column from the right, which shows penetration to be expected in the middle of the market when the cable system carries the distant signals allowed by the proposed rules. Lines (1) through (5) in Table 4 give penetration estimates for markets with successively lower levels of local service, pretty much covering the full range to be found in the top-100 markets. As the level of over-the-air service declines, expected penetration increases from 20 to 35 percent. From line (1) to line (2), this reflects the higher value put on imported signals when fewer are available locally. From line (2) to line (5), it reflects the fact that cable offers improved reception of more UHF network stations.

Line (6) shows the effect of increasing income. On all other lines, median household income is taken to be $10,000 per year, roughly the average value in the areas served by the 63 systems in my sample. For line (6), income is increased to $12,500, corresponding to an increase in real income of a little more than 2 percent per year over a 10-year period. Comparing line (6) and line (2), we note that this increase in real income is sufficient to increase expected cable penetration by 6 percentage points, from 22 to 28 percent.[26]

Similarly, line (7) shows the effect of a large reduction in the subscription price charged for cable service, such as might be possible for a subsidized or nonprofit cable system. On all other lines, price is $63 per year, the average for systems in my sample. For line (7), price is reduced to $36 per year ($3 per month). Comparing line (7) and line (2), this reduction in price is seen to increase expected penetration substantially, from 22 to 39 percent.

Line (8) shows the effect of overlapping signals that are significantly viewed in the local market and thus carried by the cable system. I assume that three network-affiliated VHF stations overlap, as shown in Figure 5. Comparing line (8) and line (2), one sees that this increases expected penetration by 3 percentage points.

It would also be of interest to investigate the effect of increased color set ownership on cable penetration. The percentage of households that have color sets has increased steadily in the past and shows little sign of leveling off yet (see Table 5).[27] Because the improved reception offered by cable is especially important for color, future increases in color set ownership will certainly tend to increase cable penetration. However, the relationship between color sets and penetration is not estimated reliably enough to justify any calculations of this effect. Consequently, I assume throughout Table 4 that half of all households own color sets, roughly the current nationwide average.

FIGURE 5

LOCATION MAP FOR EXPECTED PENETRATION EXAMPLES

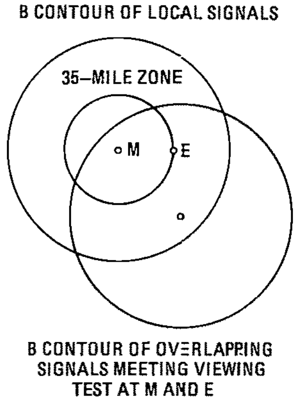

B CONTOUR OF LOCAL SIGNALS
35–MILE ZONE
M
E
B CONTOUR OF OVERLAPPING SIGNALS MEETING VIEWING TEST AT M AND E

TABLE 5

GROWTH OF COLOR SET PENETRATION

DATE	COLOR SET PENETRATION (PERCENTAGE)	RATE OF INCREASE (PERCENTAGE POINTS PER YEAR)
MARCH 1966	11	
NOVEMBER 1966	17	9.0
NOVEMBER 1968	32	7.5
FEBRUARY / MARCH 1970	43	8.5
FEBRUARY / MARCH 1971	49	6.0

So far, I have been discussing expected penetration—the figure outside parentheses in the penetration columns. But these are only *average* values for particular kinds of markets. Any individual market may experience more or less penetration, depending on factors other than those that I have been able to take into account. The figures in parentheses in Table 4 under the expected values show a likely range of penetrations for each market situation. These ranges are approximate 80-percent confidence intervals, so that there is about one chance in 10 that penetration will exceed the higher figure, and another one chance in 10 that it will fall short of the lower figure, in any particular instance. These ranges are fairly broad—generally on the order of plus or minus 10 percentage points around the expected value. Even after taking account of all of the factors included in my model, considerable uncertainty remains about how many customers a particular cable system will attract.

---

[26] This is the effect of income growth alone. Other things—cable technology and costs, public tastes, the FCC rules themselves—will also change over the next ten years in ways that are difficult to predict. Thus line (6) should not be taken to be a prediction of penetration to be expected in 1981.

[27] These figures are American Research Bureau (ARB) estimates printed in recent issues of the *Television Factbook* [10].

From examining relationships in the sample systems, we know that penetration depends on where the cable system is located: the farther away from television transmitters, the higher penetration usually is. Table 4 therefore shows expected penetration at two different locations within the market, as shown on the map (Figure 5): *M*, in the middle of the market, and *E*, at the edge at the 35-mile zone. Comparing the figures in the two columns at the right in Table 4, one sees that expected penetration at the edge of the 35-mile zone generally runs 12 or 13 percentage points higher than penetration in the middle of the market.

A final comparison answers the question, How much of a boost does cable get from the proposed rules? The middle columns in the table show the penetration that cable systems could expect without carrying any distant signals, that is, if their only selling point were improved reception of local stations. These figures generally run 5 to 10 percentage points below the corresponding figures for systems carrying all of the distant signals allowed by the proposed rules.

☐ **Expected cable penetration: three actual cases.** It is also instructive to apply the estimated relationships to predict penetration in some *actual* markets. I do so for three cities—one chosen near the top, one near the middle, and one near the bottom, of the top-100 markets. The results are shown in Table 6.

The patterns of expected penetration for St. Louis, Salt Lake City, and Evansville (Table 6) are generally similar to those for the hypothetical markets discussed above (Table 4), with one exception: There is a much greater diversity of relationships between penetration

TABLE 6

EXPECTED CABLE PENETRATION IN TOP– 100 MARKETS THREE ACTUAL CASES[a]

MARKET RANK	LOCAL SIGNALS[b]	OVERLAPPING SIGNALS ON CABLE	PENETRATION[c] (PERCENTAGE) WITHOUT DISTANT SIGNALS		DISTANT SIGNALS[b] ON CABLE	PENETRATION[c] (PERCENTAGE) WITH DISTANT SIGNALS	
			MIDDLE	EDGE		MIDDLE	EDGE
11.	ST LOUIS, MISSOURI 3NV, 1IV, 1EV	NONE	21 (14–31)	23 (16–33)	2I, 1E	26 (17–36)	28 (19–38)
49.	SALT LAKE CITY, UTAH 3NV, 2EV	NONE	19 (13–28)	22 (15–31)	3I, 1E	25 (17–36)	29 (20–40)
86.	EVANSVILLE, INDIANA 1NV, 2NU, 1EV	NONE	22 (15–31)	54 (42–66)	2I, 1E	29 (20–39)	63 (51–73)

[a] DATA ON SIGNALS RECEIVABLE OVER THE AIR, UHF AND COLOR SET PENETRATION, AND MEDIAN FAMILY INCOME ARE TAKEN FROM THE SOURCES CITED IN THE DESCRIPTION OF DATA, SEC 2. ANNUAL PRICE OF CABLE SERVICE IS ASSUMED TO BE $63

[b] N MEANS NETWORK, I MEANS INDEPENDENT, E MEANS EDUCATIONAL, V MEANS VHF, U MEANS UHF.

[c] THE TOP FIGURE IS THE MOST LIKELY VALUE. THE FIGURES IN PARENTHESES SHOW AN APPROXIMATE 80–PERCENT CONFIDENCE INTERVAL; THUS, THERE IS ABOUT 1 CHANCE IN 10 THAT PENETRATION WILL FALL SHORT OF THE LOWER FIGURE, AND 1 CHANCE IN 10 THAT IT WILL EXCEED THE HIGHER FIGURE.

in the middle and penetration at the edge of the market in the three actual cases. The edge of the market is only 2 to 4 percentage points higher than the middle for St. Louis and Salt Lake City, while the spread for Evansville is a whopping 34 percentage points. In contrast, the difference in the hypothetical examples consistently runs around 12 or 13 percentage points.

Income is the reason for the small difference in St. Louis and in Salt Lake City. Income in the central cities is considerably higher than it is toward the edges of these markets. The income effect thus tends to raise penetration in the middle compared to the edge, partially offsetting the effect of poorer over-the-air reception at the edge.

In Evansville, on the other hand, the reception effect is very much dominant. Two of Evansville's network affiliates are low-powered UHF stations. The *B* contour of one of these stations falls inside the 35-mile zone; the other does not extend much farther. Thus, primary network service over the air is quite poor at the edges of the market, and cable offers a very substantial improvement in reception of popular network programs.

As in the hypothetical cases discussed above, a 2-percent-per-year increase in real income over the next decade could be expected to increase penetration substantially, adding about 6 percentage points to each of the figures shown in Table 6.

4. Conclusion

■ By analyzing 63 existing cable television systems, all located where over-the-air television service is good, I have been able to estimate the fraction of households that can be expected to subscribe to systems operating in the top-100 markets under the FCC's new proposed rules, which allow some carriage of distant signals. In markets now served by three network-affiliated VHF stations, one can expect about 20 to 25 percent of households located in the center of the market to subscribe to cable. In markets where some of the network stations broadcast on UHF channels, somewhat higher penetration (30 to 35 percent) can be expected because of the reception problems peculiar to UHF. These figures increase by a few percentage points in high-income areas and places where stations in adjacent markets also qualify for cable carriage.

At the edges of the market, expected penetration is considerably more variable, depending heavily on the signal strength of local network stations. In markets where all three network affiliates have powerful VHF transmitters, expected penetration is only slightly higher at the edges of the market than it is in the center. But if some of the affiliates are low-powered UHF stations, poor over-the-air reception at the edges of the market can push expected penetration over 60 percent.

Distant signals carried in accordance with the FCC's proposed rules contribute substantially to expected penetration. Without distant signals, penetration would be 5 to 10 percentage points lower than the figures cited above.

Continued growth in real income and in the percentage of households with color television sets can be expected to increase cable penetration in the future. Income alone, growing at 2 percent per year, would add about 6 percentage points to the figures cited 10 years from now. I have not been able to quantify the relationship between color sets and cable penetration, but certainly the continued spread of color sets will also tend to increase penetration.

Several points of policy importance can be made as a result of this analysis. First, the expected penetration rates seem rather low (except possibly in all-UHF markets) to support the sorts of innovative programming and other services that many hope cable will provide.[28]

Second, since the estimates are based on cable systems providing traditional services only (better reception plus distant signals), new services may themselves push penetration rates higher than those estimated here. Thus we confront the familiar chicken-or-egg problem: more subscribers are needed to support new services, new services are needed to attract more subscribers.

Third, the estimates assume that subscription fees are set at the level that profit-making companies usually charge. Substantially lower fees would substantially increase penetration. (In one of the cases shown in Table 4, a reduction in annual fees from $63 to $36 increased expected penetration from 22 to 39 percent.) This suggests that one way out of the problem may be new kinds of ownership arrangements: cable systems owned by entities that are not much interested in near-term profits. If the long-term social benefits are judged to be large, such arrangements may well be desirable.

Fourth, there seems to be some danger of "cream skimming;" if cable is allowed to grow in an uncoordinated way, cable operators may well wire up communities on the fringes of the markets where higher penetration is expected and leave the central cities unserved (or at best served by their own separate, impoverished systems).

Finally, the impact on television broadcasting stations of cable systems operating under the new rules will be quite small. In earlier studies,[29] I have estimated that cable would increase the revenues of UHF independent stations by about 20 percent over the next decade, while reducing the revenues of other stations in the top-100 markets by roughly 20 percent. But these earlier studies assumed a much stronger lineup of distant signals than would be allowed under the new rules and substantially higher penetration than is predicted in this study. Thus the earlier estimates of the impact of cable need modification in two ways. First, fewer distant signals mean that the audiences of local stations are less fragmented; on this ground, revenues for all stations should tend to be higher than previously estimated. Second, lower cable penetration means that the impact of cable, both positive and negative, is reduced; on this ground, UHF independents would be helped less and the other stations hurt less than previously estimated. On balance, then, one would expect cable under the new rules to increase the revenues of independent UHF stations by about the amount estimated before, while reducing the revenues of other stations by considerably less.

Thus it appears that the new rules meet the objectives set for them by the FCC: to get cable moving without jeopardizing over-the-air broadcasting. Over-the-air broadcasting is certainly not placed in jeopardy, and the proposed distant signal allowance may well suffice to get cable moving in the cities. But more is necessary to *keep* cable moving, for the rules by themselves are probably not sufficient to make cable profitable in most of the 100 largest markets. To succeed in the cities, cable must attract customers with new services in addition to the traditional package of better reception plus distant signals.

---

[28] Recent Rand work on the costs of cable systems [2] suggests that penetrations on the order of 40 percent are necessary to support advanced systems with sophisticated facilities for local origination.

[29] Park [5,6], summarized in Park [7].

Postscript

■ As this article goes to press, the FCC has just adopted cable regulations[30] that differ in some ways from those proposed in their letter and outlined above.[31] The most important difference is that under the new regulations, syndicated programs under contract to local stations are given exclusivity protection against distant signals. For example, a cable system in New York could not carry on its distant signal channels any programs that are under contract to any New York television station. The effect is twofold: the cost of filling the distant signal channels is increased, because the cable system must switch among several distant stations to avoid carrying proscribed programs, and the attractiveness of the distant signals is decreased, because many of the best programs cannot be carried. On both counts, the probability that cable can succeed in the cities by providing traditional services is only reduced by the exclusivity provisions of the new regulations. The magnitude of this effect, which will vary from market to market for a number of reasons, is the subject of a current study at Rand.

References

1. Comanor, W. S. and Mitchell, B. M. "Cable Television and the Impact of Regulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 154–212.
2. Johnson, L. L. et al. "Cable Communications in the Dayton Miami Valley: Basic Report." Rand Report R-943-KF/FF, January 1972.
3. Lydon, C. "New Rules on Cable TV Limit Growth in Cities." *New York Times*, February 4, 1972, p. 1.
4. McGowan, J. J., Noll, R. G., and Peck, M. J. "Prospects and Policies for CATV." Mimeographed. Washington, D. C.: The Brookings Institution, March 1971.
5. Park, R. E. "Cable Television and UHF Broadcasting." Rand Report R-689-MF, January 1971, forthcoming as "Cable Television, UHF Broadcasting, and FCC Regulatory Policy," in *The Journal of Law and Economics*, Vol. 15, No. 2 (October 1972).
6. ———. "Future Growth of Cable Television." *Journal of Broadcasting*, Vol. 15, No. 3 (Summer 1971) pp. 253–64, also appears in R. E. Park, "Potential Impact of Cable Growth on Television Broadcasting," Rand Report R-587-FF, October 1970, pp. 8–27.
7. ———. "The Growth of Cable TV and Its Probable Impact on Over-the-Air Broadcasting." *The American Economic Review*, Vol. 61, No. 2 (May 1971), pp. 69–73.
8. ———. "Prospects for Cable in the 100 Largest Television Markets." Rand Report R-875-MF, October 1971.

---

[30] See Lydon [3].

[31] Page 143 *infra*.

9. Television Digest, Inc. *CATV and Station Coverage Atlas*. Washington, D. C.: Television Digest, Inc., appropriate years.
10. ———. *Television Factbook*. Services Volume. Washington, D. C.: Television Digest, Inc., annual.
11. U. S. Bureau of the Census. *County and City Data Book 1967*. Washington, D. C.: U. S. Government Printing Office, 1967.
12. ———. *Current Population Reports*. Series P-60, No. 75, "Income in 1969 of Families and Persons in the United States," Washington, D. C.: U. S. Government Printing Office, 1970.
13. U. S. Federal Communications Commission. Letter from Dean Burch, Chairman, to the Chairmen of the Senate Communications Subcommittee and the House Communications and Power Subcommittee, August 5, 1971.

Corporate investment under uncertainty and Pareto optimality in the capital markets

Michael C. Jensen

Associate Professor
Graduate School of Management
University of Rochester

and

John B. Long, Jr.

Assistant Professor
Graduate School of Management
University of Rochester

Within the context of a mean-variance equilibrium model of the pricing of capital assets, this paper investigates the allocation of investment in new risky opportunities which results from the collective behavior of firms, each of which attempts to maximize the net increase in its market value. This allocation is then compared to those allocations which (1) maximize nominal social wealth or (2) maximize social welfare. These comparisons are made under a variety of assumptions concerning the nature of the new opportunities, investor attitudes toward risk, and the number and characteristics of firms in the economy. With the exception of certain special cases, it is found that the private allocation of investment does not correspond to that which either maximizes social wealth or maximizes social welfare.

1. Introduction

■ Irving Fisher demonstrated that in a world of perfect certainty and perfect capital markets[1] an individual will maximize his utility of consumption by organizing his production so as to maximize his wealth (as measured by the discounted present value of his income stream). Given this wealth maximizing income stream the individual

---

Michael C. Jensen earned the A.B. degree in economics at Macalester College (1962) and both the M.B.A. and the Ph.D. in economics at the University of Chicago, Graduate School of Business (1964 and 1968, respectively).

John B. Long, Jr. received the B.A. degree in mathematics from Rice University (1966) and the Ph.D. in industrial administration from Carnegie-Mellon University (1971).

The authors wish to thank Fischer Black, Eugene Fama, and Martin Bailey for benefits obtained from conversations on these issues. The authors retain responsibility for any errors which may remain in the analysis. This research was partially supported by a National Science Foundation grant to M. C. Jensen.

Subsequent to the completion of the first draft of this paper (originally entitled "On the Inconsistency Between 'Optimal' Corporate Investment and Pareto Optimality") in September, 1969, others reported results similar to these. See J. Stiglitz, "On the Optimality of the Stock Market Allocation of Investment," presented at the Far Eastern Meetings of the Econometric Society, Tokyo, June 1970, forthcoming in the *Quarterly Journal of Economics*, and E. Fama and A. Laffer, "Perfect Competition and the Market Value Rule in the Two-Parameter Model," mimeographed, University of Chicago, October, 1971, forthcoming in *The Bell Journal of Economics and Management Science*.

[1] Perfect in the sense that no individual can influence the price of any firm or asset by his purchases or sales, transactions costs are zero, and borrowing and lending rates are identical for all individuals. See [4].

can then achieve an optimal time pattern of consumption by borrowing or lending in the capital markets. Hirschleifer and Fama and Miller[2] demonstrate that Fisher's approach can be used to show that under perfect certainty and perfect capital markets a firm which chooses its investment strategy so as to maximize the value of the firm will provide maximum utility of consumption to its stockholders. These are extremely appealing results for those concerned with the problem of formulating investment criteria for a firm. One would hope that similar results would apply in the context of uncertainty. A large number of authors, including Modigliani and Miller, Fama, Lintner, and Mossin,[3] have postulated value maximization as the appropriate investment criterion for a firm under uncertainty as well as certainty. Many textbooks in finance such as those by Weston and Brigham and Van Horne[4] also postulate this criterion.

The purpose of the paper is twofold: (1) given the general equilibrium model of the pricing of capital assets developed by Sharpe, Lintner, Fama, and Mossin,[5] to derive the allocation of investment to new opportunities which occurs when each firm attempts to maximize the net increment in its market value, and (2) to derive those investment allocations which (a) maximize social welfare and (b) maximize nominal social wealth.

Long[6] has shown that the investment decision made by a given firm attempting to maximize the net increase in its market value may not be optimal from the standpoint of its stockholders even if that decision maximizes their nominal wealth. In this paper we show, in essence, that when one takes account of the reaction of other firms to a given firm's decisions and of the interrelations among the market values of the various firms, the private allocation in the economy of investment in new opportunities by value maximizing firms will not, except in some special cases, maximize either nominal social wealth or social welfare.

In Section 2 we consider a simple formulation of the model under the assumption that the price of risk expressed as the ratio of total mean excess returns to total variance is constant. In Section 3 we consider a more general formulation of this model taking account of the interaction of individual firm investment decisions. In Section 4 we consider the formulation of the model under the assumption that the price of risk expressed as the ratio of total mean excess returns to total standard deviation is constant. Section 5 contains a brief summary of our results and our conclusions.

2. Social and corporate investment criteria—a simple case

■ **The valuation model.** We shall concern ourselves here with a single-period mean-variance model of the pricing of capital assets under uncertainty such as that developed by Sharpe, Lintner, Fama, and Mossin.[7] Consider a single-period world in which all investors are risk-averse expected utility maximizers whose consumption-investment decisions can be characterized by the maximization of the preference function $G_i(c_i, e_i, v_i)$, where c_i is the dollar amount of the ith investor's consumption at the beginning of the period, e_i is the expected value of the cash flow to be generated one period hence by the ith investor's portfolio, and v_i is the variance of this cash flow. We assume $\partial G_i/\partial c_i > 0$, $\partial G_i/\partial c_i^2 < 0$, $\partial G_i/\partial e_i > 0$, $\partial G_i/\partial v_i < 0$, all assets are infinitely divisible, all investors have homogeneous expectations, transactions costs and taxes are zero, and the prices of consumption goods are fixed. Under these assumptions the equilibrium value of the jth firm will be given by:

---

[2] In [5] and [3], respectively.

[3] See [13], [2], [7], and [15], respectively.

[4] See [18] and [17].

[5] In [16], [7,8], [1,2], and [14], respectively.

[6] In [9].

[7] See note 5 *supra.*

$$V_j = \frac{1}{r}[\bar{D}_j - \lambda\sigma_{jM}] \quad \text{for all } j \tag{1}$$

where

$\bar{D}_j = E(\tilde{D}_j)$ = the expected value of the total end-of-period cash flow to the stockholders of firm j,

$r = 1 + i$, where i is the one-period riskless rate of interest at which every consumer can borrow or lend,[8]

$\sigma_{jM} = \sum_k \sigma_{jk}$ = covariance of the total cash flows of the firm, $\tilde{D}_j$, with $\tilde{D}_M$ the total cash flows from all firms,

$$\sigma_{jk} = \begin{cases} \operatorname{var}(\tilde{D}_j), & j = k \\ \operatorname{cov}(\tilde{D}_j, \tilde{D}_k), & j \neq k, \end{cases}$$

σ_M^2 = variance of $\tilde{D}_M$,

$\lambda = [\bar{D}_M - rV_M]/\sigma_M^2$ = market price per unit of risk, and

$V_M = \sum_k V_k$ = total market value of all firms.

Thus in equilibrium the value of the jth firm is the present value (discounted at the risk-free rate) of the certainty equivalent of the random payment $\tilde{D}_j$. The certainty equivalent is simply the expected payment $\bar{D}_j$ minus a risk discount given by the product of λ, the price per unit of risk, and the risk of the firm, given by the sum of its variance and its covariances with all other firms. We shall also find it convenient, as is often done, to write the valuation equation as

$$V_j = \frac{1}{r}\left[\bar{D}_j - \theta\frac{\sigma_{jM}}{\sigma_M}\right], \tag{2}$$

where

$$\theta = [\bar{D}_M - rV_M]/\sigma_M.$$

☐ **Consumer welfare and wealth maximization.** In dealing with the effects of corporate investment, changes in consumer wealth are often taken as a surrogate for changes in consumer welfare. As pointed out by Long,[9] however, this approach is not always correct. As one might suspect, cases in which maximum wealth does not correspond to maximum welfare generally involve situations where capital investment influences not only the wealth of consumers but also the rates of return on assets available to consumers in the market.

---

[8] The model as formulated here can also be interpreted as an infinite time horizon model where all probability distributions of single-period cash flows are assumed stationary and r is the constant single-period riskless rate of interest.

[9] In [9].

In these situations changes in nominal consumer wealth due to corporate investment may not correspond in direction to the concommitant changes in consumer utility.

In light of the possible divergence of wealth and welfare, the effects of corporate investment on these variables will be individually analyzed. Specifically, we will compare the aggregate amount and distribution of corporate investment resulting from the behavior of value maximizing firms with the allocations of investment which (1) maximize consumer welfare and (2) maximize consumer wealth.

These comparisons will be made under two alternative assumptions regarding the behavior of the market trade-off between risk and return. Such assumptions must be made if we are to avoid a detailed specification of each consumer's preference function. As shown in Section 4, however, the two alternatives used here do not yield substantially different results. In the remainder of Section 2 and in Section 3, λ, as defined above, will be held constant. As shown by Lintner,[10] this corresponds to the case in which all consumer preference functions exhibit constant absolute risk aversion. In Section 4, θ is held constant. This case is of interest since, as will be shown, it is one in which the welfare maximizing allocation of investment is identical to the wealth maximizing allocation. Unlike the case in which λ is held constant, however, there appears to be no plausible preference structure which will guarantee the constancy of θ.

☐ The social investment criteria.

The welfare maximizing criterion

In order to develop the social and corporate investment criteria and to illustrate their characteristics let us first consider a simple case. (We present a more general formulation of the problem and investigate the characteristics of its solution in Section 3 below.)

Let us consider a new investment opportunity[11] which promises a random return per dollar invested, $\tilde{\rho}$, with

$$E(\tilde{\rho}) = \bar{\rho},$$

$$\mathrm{var}(\tilde{\rho}) = \sigma_{\rho}{}^2,$$

and

$$\mathrm{cov}(\tilde{\rho},\tilde{D}_j) = \sigma_{j\rho}.$$

We assume that the joint probability distribution of $\tilde{D}_1$, $\tilde{D}_2$, $\cdots$ $\tilde{D}_N$ and $\tilde{\rho}$ is independent of which firm takes the project. Denote the amount invested in this project by firm j to be I_j and assume that the stochastic characteristics of $\tilde{\rho}$ are independent of the amount invested, I_j; that is, there are constant stochastic returns to scale. Our problem here is to find the optimal total investment in this project from society's viewpoint using as our criteria the maximization of the welfare of every investor in the economy. Our analysis is partial equilibrium to the extent that we assume all product prices in the consumption markets are fixed and we concentrate on the capital markets.

---

[10] See [6].

[11] New in the sense that there has been no prior investment in this opportunity by any firm.

To determine the socially optimum investment in the new opportunity we must refer back to the structure of market equilibrium. In the mean-variance model, the typical consumer will find it optimal at equilibrium to hold an investment portfolio consisting of some quantity of the riskless asset together with a fraction of the "market portfolio" (that portfolio which contains all risky shares). To be specific, consider a given investor and let him hold a fraction α of the market portfolio at equilibrium. Suppose now that the investor increases α, borrowing to finance the added investment in the market portfolio. The rate of increase in the investor's expected end-of-period wealth as α is increased is then $\frac{\partial e}{\partial \alpha} = (\bar{D}_M - rV_M)$. The rate of increase in the variance of his ending wealth is $\frac{\partial v}{\partial \alpha} = 2\alpha\sigma_M{}^2$. Thus the rate of exchange of expected ending wealth for variance of ending wealth is, for this investor, given by

$$\frac{\bar{D}_M - rV_M}{2\alpha\sigma_M{}^2} = \frac{\lambda}{2\alpha}. \tag{3a}$$

Suppose now that a firm, say firm j, increases its investment, I_j, in the new opportunity described above. The investor being considered owns a fraction of α of firm j and thus receives that fraction of the returns from the new investment. He must also share a fraction α of the cost of the new investment. Assuming then that he borrows to finance his share of the investment, the rate of increase in his expected ending wealth as I_j increases is $\frac{\partial e}{\partial I_j} = \alpha(\bar{p} - r)$ and the rate of increase in the variance of his ending wealth is $\frac{\partial v}{\partial I_j} = 2\alpha^2\sigma_{Mp}$, where σ_{Mp} is the covariance between $\bar{p}$ and the dollar returns from all assets in the market. The rate of exchange of expected ending wealth for variance of ending wealth offered by the investment is then

$$\frac{\alpha(\bar{p} - r)}{2\alpha^2\sigma_{Mp}} = \left(\frac{1}{2\alpha}\right)\left(\frac{\bar{p} - r}{\sigma_{Mp}}\right). \tag{3b}$$

The investor will gain from increased investment in the new opportunity as long as the rate of exchange given by (3b) is more favorable than that given by (3a) or, equivalently, as long as

$$\bar{p} > r + \lambda\sigma_{Mp}. \tag{3c}$$

Note that this conclusion is independent of α and hence is true for all investors. The right-hand side of (3b) is a function of I_j since, as investment is made in the new opportunity, the returns from this investment become a part of the total returns from the market portfolio. The socially optimum level of investment is attained when the risk-return trade-off offered by the investment equals that available to consumers in the market. At this point the inequality given by (3c) becomes an equality. Thus if it is socially desirable to make some positive investment in the new opportunity the optimum level of investment will satisfy

$$\bar{p} = r + \lambda\sigma_{M'p}, \tag{4}$$

where $\sigma_{M'\rho}$ is the covariance between $\bar{\rho}$ and the *post*-investment return of the market portfolio.

There are two clarifying points that should be made concerning the criterion given by (4). First, it may not be socially desirable to make any positive investment in the new opportunity. This would be the case where, even at a zero level of investment, the risk-return trade-off offered by the investment is unfavorable, i.e., where $\bar{\rho} < r + \lambda\sigma_{M'\rho}$ at $I_j = 0$. Thus a more general criterion is that $\bar{\rho} \leq r + \lambda\sigma_{M'\rho}$ at the optimum level of investment with strict equality holding when the optimum level is positive. The second point to be made is that when several firms are simultaneously investing, (4) must be simultaneously satisfied for each investing firm. This point becomes important when we consider the possibility of differences among the new opportunities of different firms.

For present purposes let us assume that the project may only be taken by a single firm j, which invests an amount I_j. In addition let us assume that no other investment or disinvestment occurs in the economy. The post-investment return of the market portfolio will then be $\tilde{D}_{M'} = \tilde{D}_M + I_j\tilde{\rho}$ and $\sigma_{M'\rho} = \sigma_{M\rho} + I_j\sigma_\rho^2$. Substitution of this into (4) then yields the welfare maximizing level of investment, I_w^*, given by

$$I_w^* = \max\left\{0, \frac{1}{\lambda\sigma_\rho^2}[(\bar{\rho} - r) - \lambda\sigma_{M\rho}]\right\}. \tag{5}$$

Inspection of (5) shows that the results are intuitively pleasing. The optimum size investment in the project is inversely related to the variance of the project returns and the degree of covariance of the project returns with all other assets in the economy and positively related to the difference between the expected return, $\bar{\rho}$, and the riskless return, r. We shall return to a consideration of (5) after deriving the social wealth maximizing criterion and the corporate investment criterion.

The social wealth maximizing criterion

The assumptions of the valuation model ensure that the portfolios of risky assets held by all investors are identical except for scale (i.e., total dollars invested), and since equilibrium requires that all assets be held, each investor's risky asset portfolio contains a given fraction of every firm in existence[12] and hence gives him a fractional claim on $\tilde{D}_M$, the total cash flow of all firms. Under these conditions the wealth of each investor increases with new investment by individual firms in a project if and only if the sum of the market values of all firms increases by an amount greater than the total amount invested. By (1) we know that

$$V_M = \sum_k V_k = \frac{1}{r}[\bar{D}_M - \lambda\sigma_M^2], \tag{6}$$

and using the definition of $\tilde{D}_{M'}$ the new aggregate value of all firms V_M is

$$V_M' = \frac{1}{r}[(\bar{D}_M + I_j\bar{\rho}) - \lambda(\sigma_M^2 + 2I_j\sigma_{M\rho} + I_j^2\sigma_\rho^2)]. \tag{7}$$

---

[12] See Mossin [14].

Let $\Delta V_M(I) = V_M' - V_M$. Subtracting I from $\Delta V_M(I)$ gives the net change in the value of all assets as

$$\Delta V_M(I) - I_j = \frac{I_j}{r}\left[(\bar{\rho} - r) - \lambda(2\sigma_{M\rho} + I_j\sigma_\rho^2)\right], \tag{8}$$

and if (8) is positive, the wealth of all investors is larger than before the new investment. Since the value of equation (8) is independent of which firm makes the investment, each investor is concerned only with the size of the investment and is indifferent as to which firm takes it.[13] Hence we drop the subscript j on I_j.

Since equation (8) is quadratic in I, there is a finite optimum scale of investment as viewed by every investor even though we have assumed constant returns to scale ($\bar{\rho}$ and σ_ρ^2 independent of I). Maximizing (8) with respect to I (subject to $I \geq 0$), provides the social wealth maximizing investment, I_s^*, as

$$I_s^* = \max\left\{0, \frac{1}{2\lambda\sigma_\rho^2}\left[(\bar{\rho} - r) - 2\lambda\sigma_{M\rho}\right]\right\}. \tag{9}$$

Comparison of (9), the social wealth maximizing investment, with (5), the social welfare maximizing investment, illustrates the inconsistency between the criteria. The level of investment by firm j which would maximize social wealth is less than the level of investment which would result in a maximum increase in utility for all investors. Indeed, for a prospective project whose returns are uncorrelated with all other cash flows (i.e., $\sigma_{M\rho} = 0$), $I_w^* = 2I_s^*$, and for this case when $I_j = I_w^*$, $\Delta V_M - I_j = 0$, the social wealth remains unchanged.

☐ **The corporate investment criterion.** Let us postulate the firm's goal to be the maximization of its market value (net of investment, of course). Given this goal, we wish to find the optimum level of investment for a particular firm in the project outlined above. For simplicity we continue to assume that no other investment or disinvestment occurs in the economy. (We shall relax this assumption in Section 3 below and consider the simultaneous solution of the investment decisions of all firms having access to the project while taking all interaction effects into account.) If the firm obtains new capital, I_j, to take the project,[14] its value, V_j', after the new market equilibrium has been reached will be

---

[13] This indifference would not exist, of course, if the joint probability distribution of $\tilde{\rho}$ and $\tilde{D}_M$ were not independent of which firm takes the project.

[14] In our discussion of the effects of new investment we have assumed that the firm has only one class of securities, common stock, and that new capital is raised through the sale of additional stock. This assumption is purely for convenience, however, since the alternative approach (in which firms are allowed to borrow at the riskless rate) yields investment rules for society and for individual firms which are identical to those derived under our approach. Note that this additional capital is obtained by net new borrowings at the rate r by investors (since the assumptions of the valuation model imply that the supply schedule of risk-free funds is perfectly elastic to the system at the rate r). The system can be closed by assuming the existence of a riskless investment opportunity (yielding constant returns to scale r) available to everyone. Then as long as the return r is sufficiently high to assure positive investment in this riskless activity the expansion (or contraction) of investments in risky ventures will take place through the contraction (or expansion) of the scale of activity in the riskless activity.

$$V_j' = \frac{1}{r}[(\bar{D}_j + I_j\bar{\rho}) - \lambda(\sigma_{Mj} + I_j\sigma_{Mp} + I_j\sigma_{jp} + I_j^2\sigma_p^2)] , \quad (10)$$

where the post-project risk of the firm is given by $\mathrm{cov}(\tilde{D}_j + I_j\tilde{\rho}, \tilde{D}_M + I_j\tilde{\rho})$, the covariance between the post-project cash flows of the firm and the market.

Now define $\Delta V_j(I) = V_j' - V_j$ as the change in the value of firm j brought about by investment of an amount I in this project. After subtracting the cost of the project, the net increase in the value of firm j is given by

$$\Delta V_j(I) - I_j = \frac{I_j}{r}[(\bar{\rho} - r) - \lambda(\sigma_{Mp} + \sigma_{jp} + I_j\sigma_p^2)] , \quad (11)$$

and if this quantity is greater than zero for some value of I_j, the project is desirable from the firm's viewpoint. Equation (11) is also quadratic in I_j, and we find the optimum value of I_j by maximizing (11) subject to the constraint that $I_j \geq 0$. The solution, I_j^*, is

$$I_j^* = \max \left\{ 0, \frac{1}{2\lambda\sigma_p^2}[(\bar{\rho} - r) - \lambda(\sigma_{Mp} + \sigma_{jp})] \right\} . \quad (12)$$

Note that the optimum size investment in the project when viewed by the jth firm involves the variance of the per-dollar returns on the project, σ_p^2, the sum of the covariances of the returns on the project with all *other* firms, and twice the covariance of the returns with firm j (since σ_{jp} is one of the terms in the first summation). Thus, even though we have assumed that the stochastic properties of the returns on this project are independent of which firm takes the project, different firms may require different expected returns on the project for acceptance.[15] All terms in (12), except σ_{jp} (the covariance of the project returns with firm j's cash flow), are identical for all firms. Thus the optimum investment in the project for firms i and j will not be identical unless $\sigma_{jp} = \sigma_{ip}$.[16] Indeed it is possible for I_j^* to be positive while I_i^* is zero if σ_{ip} is sufficiently greater than σ_{jp}.

This differential demand for the project is surprising, since one might intuitively believe that in a perfect capital market with homogeneous expectations the optimum size investment in this type of project should not depend on which firm takes the project. The result is due to the fact that if the project is accepted the values of all firms in the economy may change as the new set of equilibrium prices is reached, and the jth firm internalizes only a portion of this total adjustment.

In addition this differential demand for the project by different firms is even more surprising in light of the fact that all investors hold portfolios which give them a fractional claim on $\tilde{D}_M$. Thus, given that the project is going to be taken (in a given amount), all investors are completely indifferent as to which firm takes it. Hence, it appears that, contrary to a common interpretation of the results of the asset pricing model outlined above, something akin to a diversification consideration will enter into the firm's choice among investment projects.

---

[15] Lintner and Mossin (in [7] and [15], respectively) noted this peculiarity in investment criteria they derived for value maximizing firms.

[16] Except in the case in which both firms might find the project unprofitable and $I_j^* = I_i^* = 0$, and $\sigma_{jp} \neq \sigma_{ip}$.

Although we have shown that the value of a *prospective* project may be different for different firms even though the probability distributions of cash flows on the project are identical for all firms, this does not imply that the total value of a given set of existing assets can be influenced by their particular groupings into firms. That is, the merger of two firms, or the purchase by one firm (at the market price) of a set of assets from a second firm will leave the total value of the two firms unchanged.[17] This is easily seen by examining the value equation (1) and noting that both the mean and covariance are additive.

We now turn to a detailed examination of the implications of the differences between the social investment criteria given by (5) and (9) and the corporate criterion given by (12).

☐ Optimum social investment vs. "optimum" corporate behavior—a first look.

Some general conclusions

Comparison of (5) with (12) reveals an inconsistency between the behavior of a value maximizing firm and the preferences of all investors regarding the level of investment by the firm. The variance of the returns on the project has twice the impact on the firm's investment, $I_j{}^*$, as it does on the social optimum investment, $I_w{}^*$, and will in general cause the firm to invest less than $I_w{}^*$. In addition the individual firm also views the risk of the project differently since its optimum investment depends on σ_{jp} in addition to σ_{Mp}, and hence if the returns of the project are positively correlated with the firm's previous cash flows this factor will cause it to invest still less than $I_w{}^*$.

In addition, comparison of (9) and (12) indicates that the investment $I_j{}^*$ which maximizes the value of the individual firm is in general inconsistent with the investment $I_s{}^*$ which maximizes the wealth of all stockholders. Thus, not only is stockholder wealth maximization inconsistent (under the assumption of constant λ) with stockholder utility maximization, but value maximization by individual firms is inconsistent with stockholder wealth maximization.

Examination of (9) and (12) indicates that under our assumptions the divergence of the social wealth maximizing criterion and the private value maximizing investment criterion can occur only when the cash flows of a project under consideration by any firm are correlated with the cash flows of other firms. If the project returns are uncorrelated with the pre-investment cash flows of all other firms the private and social wealth maximizing criteria give identical investments.

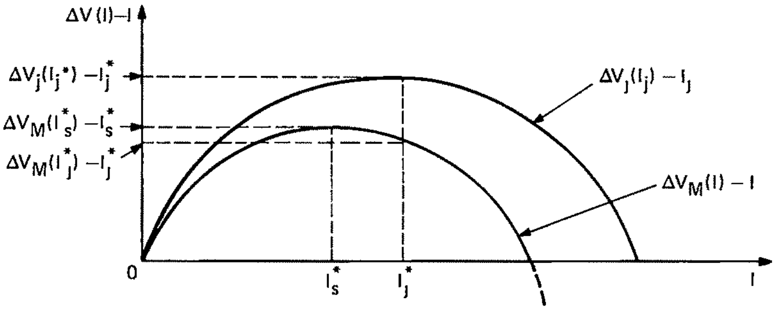
Figure 1 illustrates the inconsistency between the corporate and social wealth maximizing investment criteria for a desirable project for which $\sigma_{jp} > 0$ and $\sigma_{Mp} - \sigma_{jp} > 0$; that is, the cash flows of the project are positively correlated with the cash flows of the existing assets of firm j and with the total cash flow from the assets of all

---

[17] Ignoring, of course, any economies of scale or monopoly elements which might arise.

FIGURE 1

NET CHANGE IN THE VALUE OF THE jth FIRM AND THE TOTAL VALUE OF ALL FIRMS DUE TO THE INVESTMENT BY FIRM j IN A DESIRABLE PROJECT FOR WHICH $\sigma_{j\rho} > 0$ AND $\sigma_{M\rho} - \sigma_{j\rho} > 0$



other firms. Under these conditions we see that the firm acting so as to maximize the net change in its value will overinvest in the sense that all investors would be wealthier if it invested only $I_s^* < I_j^*$. The reason that the aggregate curve lies everywhere below the curve for firm j in Figure 1 is the fact that although firm j receives all of the benefits from the increased mean returns due to its investment, it does not internalize all of the risk associated with the investment. The acceptance of the project by firm j serves to increase the riskiness of any firm k for which $\sigma_{k\rho}$ is positive, and thus its value in the new equilibrium will be lower. The reverse, of course, is true if $\sigma_{k\rho}$ is negative. (If the cash flows of the project are on balance negatively correlated with the cash flows of all other firms, i.e., $\sigma_{M\rho} - \sigma_{j\rho} < 0$, the two curves in Figure 1 are reversed and there will be underinvestment in the project as reviewed by every investor.) Note, however, that the utility maximizing investment I_w^* is even greater than I_j^*, and therefore there will be systematic underinvestment from the viewpoint of utility maximizing investors.

3. Equilibrium investment with a tatonnement market clearing process

■ **A more general model of the characteristics of investments.** Thus far we have considered only the case where a single firm had the opportunity to take a project and no investment or disinvestment by other firms was allowed. We consider here the situation where more than one firm has access to the project, and thus each firm in making its investment decision must take into account the actions of all other firms. We continue to assume that $\tilde{\rho}_j$, ($j = 1, 2, \ldots, N$) is a set of identically distributed random variables. In addition we must now consider the covariance between the returns on the project when taken by two different firms. We shall assume that the degree of dependence of the returns on the new project among firms is characterized by a correlation coefficient, β, ($0 \leq \beta \leq 1$), such that

$$\text{cov}(\tilde{\rho}_j, \tilde{\rho}_k) = \beta\sigma_\rho^2 \qquad \text{for all } j \text{ and } k. \qquad (13)$$

Note that we are constraining the correlation of project returns among firms to be identical. In addition it seems unlikely that the correlation of returns on a given project taken by two different firms would be negative.

☐ **The corporate investment criterion.** If firm j invests I_j dollars in the project while other firms in the market invest amounts $\{I_k\}$ $k \neq i$, then the post-investment cash flows for firm j and for the market are

$$\tilde{D}_j' = \tilde{D}_j + I_j\tilde{\rho}_j\,, \tag{14}$$

$$\tilde{D}_M' = \tilde{D}_M + \sum_k I_k\tilde{\rho}_k\,, \tag{15}$$

and using (13) we find the post-investment risk of firm j to be

$$\begin{aligned}\operatorname{cov}(\tilde{D}_j', \tilde{D}_M') &= \operatorname{cov}(\tilde{D}_j, \tilde{D}_M) + \operatorname{cov}(\tilde{D}_j, \sum_k I_k\tilde{\rho}_k) + \operatorname{cov}(I_j\tilde{\rho}_j, \tilde{D}_M)\\ &\qquad + \operatorname{cov}(I_j\tilde{\rho}_j, \sum_k I_k\tilde{\rho}_k)\\ &= \sigma_{jM} + \sigma_{j\rho}\sum_k I_k + I_j\sigma_{M\rho}\\ &\qquad + I_j\sum_{k\neq j} I_k\beta\sigma_\rho^2 + I_j^2\sigma_\rho^2\,.\end{aligned} \tag{16}$$

Using (16) and (1) we find the net change in the total market value of firm j to be

$$\begin{aligned}&\Delta V_j(I_j) - I_j\\ &\quad = \frac{1}{r}\,[I_j(\bar{\rho} - r) - \lambda(\sigma_{j\rho}I + I_j\sigma_{M\rho} + I_j\sum_{k\neq j} I_k\beta\sigma_\rho^2 + I_j^2\sigma_\rho^2)]\,,\end{aligned} \tag{17}$$

where $I = \sum_k I_k$ is the total aggregate investment. Note that the form of (17) indicates that under our assumptions the net change in the value of firm j, for any given investment, I_j, is independent of the distribution of the investment by all other firms and (if $\beta > 0$) depends only on the total investment and not its distribution among firms. Maximizing (17), the net change in the value of the firm, with respect to I_j subject to the constraint that $I_j \geq 0$ (and given the total amount invested by other firms, $\sum_{k\neq j} I_k$), we find firm j's optimum investment

$$\begin{aligned}I_j^* = \max\Bigg\{0, \left(\frac{1}{2\lambda\sigma_\rho^2}\right)&\\ \times [(\bar{\rho} - r) - \lambda(\sigma_{M\rho} + \sigma_{j\rho} + \beta\sigma_\rho^2 &\sum_{k\neq j} I_k)]\Bigg\}\,.\end{aligned} \tag{18}$$

If $\beta > 0$ and if other firms also invest in the project, we see by comparison of (18) with (12) that the riskiness of the project as perceived by firm j increases, and its optimum investment, I_j^*, is lower.

☐ **Competitive equilibrium and aggregate private investment.** Given a tatonnement market clearing process, the equilibrium investment in our hypothetical project for all firms will occur when

$$\begin{aligned}I_j^* = \max\left\{0, \left(\frac{1}{2\lambda\sigma_\rho^2}\right)[(\bar{\rho} - r) - \lambda(\sigma_{M\rho} + \sigma_{j\rho} + \beta\sigma_\rho^2 \sum_{k\neq j} I_k^*)]\right\},&\\ j = 1, 2, \ldots, N\,.&\end{aligned} \tag{19}$$

That is, equilibrium will occur when all firms are simultaneously at their optimum levels of investment.

To simplify the exposition of the solution to the set of relations (19), let

$$a_j = \left(\frac{1}{2\lambda\sigma_\rho^2}\right)[(\bar{\rho} - r) - \lambda(\sigma_{M\rho} + \sigma_{j\rho})], \quad j = 1, 2, \ldots, N. \quad (20)$$

The equilibrium conditions (19) can then be written as:

$$I_j^* = \max\left\{0, a_j - \frac{\beta}{2}\sum_{k \neq j} I_k^*\right\}, \quad j = 1, 2, \ldots, N. \quad (21)$$

It is clear from (21) that if $a_j \leq 0$, then $I_j^* = 0$. Restricting our attention then to those firms for which $a_j > 0$, let us introduce the convention that the firms in this restricted set are indexed such that $0 < a_1 \leq a_2 \leq \cdots \leq a_n$, $(n \leq N)$.

The general case

In the general case where $0 < \beta \leq 1$ and where the a_j's are not identical the solution can be characterized by the following statement: If, at equilibrium, $I_j^* > 0$ and $I_k^* = 0$, then $a_j > a_k$, or on the basis of our indexing convention, $j > k$ where $1 \leq k < j \leq n$. The cutoff index point and the specific solution are given by the following theorem.

THEOREM: Among the set of firms $j = 1, 2, \ldots, N$, the equilibrium distribution of investment is specified by:

$$I_j^* = 0, \quad j < k \quad (\text{if } k = 1, I_j^* > 0, \quad j = 1, 2, \ldots, N), \quad (22a)$$

$$I_j^* = \frac{[4 + 2(n - k)\beta]a_j - 2\beta\sum_{t=k}^{n} a_t}{(2 - \beta)[2 + (n - k)\beta]}, \quad j \geq k, \quad (22b)$$

where k is the smallest index number $(1 \leq k \leq n)$ for which

$$a_k > \frac{\beta\sum_{j=k}^{n} a_j}{2 + (n - k)\beta}. \quad (23)$$

The existence of this solution is guaranteed and it is unique. (Existence and uniqueness depend on the fact that $0 \leq \beta \leq 1$.)

PROOF: The equilibrium condition is given by (21). If we let $S = \sum_{k=1}^{n} I_k^*$, then (21) can be expressed as

$$I_j^* = \max\{0, a_j - (\beta/2)(S - I_j^*)\}, \quad j = 1, 2, \ldots, N, \quad (24)$$

or

$$I_j^* = \max\{0, [1 - (\beta/2)]^{-1}[a_j - (\beta/2)S]\}, \quad j = 1, 2, \ldots, N. \quad (25)$$

From (25), we can make the following statement: If $I_j^* > 0$ and if $a_k \geq a_j$, then $I_k^* \geq I_j^* > 0$. Thus, since $a_1 \leq a_2 \leq \cdots \leq a_n$, the set of firms which invest positive amounts at equilibrium can be characterized by a cutoff index number, k, such that $I_j^* = 0, j < k$, and $0 < I_k^* < I_{k+1}^* \leq \cdots \leq I_n^*$.

Using the above characterization of the solution, S can be evaluated from (25) as:

$$\begin{aligned} S &= \sum_{j=1}^{n} I_j^* = \sum_{j=k}^{n} I_j^* \\ &= [1 - (\beta/2)]^{-1}[\sum_{j=k}^{n} a_j - (n - k + 1)(\beta/2)S] \qquad (26) \end{aligned}$$

or

$$S = \frac{2 \sum_{j=k}^{n} a_j}{2 + (n - k)\beta}, \qquad (27)$$

where [from (25)] k is the smallest index number for which $a_k > (\beta/2)S$, or equivalently, for which (23) is satisfied.

Finally, given the definition of k, substitution from (27) into (25) yields the theorem given by (22a) and (22b).

☐ **The social welfare maximizing criterion.** Given that the new investment project outlined above is available to some (or all) firms in the economy, our goal here is to answer the following question: If investors could dictate the investment strategy of all firms in the market, what would their investment policy be? Using the necessary conditions for the maximization of each investor's utility given in (4), we have for the current case

$$(\bar{\rho} - r) - \lambda(\sigma_{M\rho} - \beta\sigma_\rho^2 \sum_{k \neq j} I_k + I_j\sigma_\rho^2) \leq 0\,, \quad j = 1, 2, \ldots, N\,. \qquad (28)$$

Using the fact that $I = \sum I_j$ and solving (28) for I_j we have (for $\beta_j < 1$)

$$I_j = \frac{1}{\lambda\sigma_\rho^2(1 - \beta)}[\bar{\rho} - r - \lambda(\sigma_{M\rho} + I\beta\sigma_\rho^2)]\,, \quad j = 1, 2, \ldots, N\,, \qquad (29)$$

which is a constant (call it Q) independent of j. Thus

$$\sum_j I_j = NQ = I\,,$$

and hence

$$I_j = Q = \frac{I}{N}, \quad j = 1, 2, \ldots, N \qquad (30)$$

is the optimum distribution of the investment across firms. That is, given $\beta < 1$ and that some positive total investment is desirable, investors would prefer that total investment be divided equally among all firms. If $\beta = 1$ they are indifferent as to its distribution among firms. Given our assumptions, these results are reasonable since allocating the investment equally among firms maximizes the possible diversification as long as the correlation, β, is less than unity.[18]

---

[18] If this intercorrelation of project returns were not identical for all firms this result would not hold.

Given the solution for $\{I_j\}$, substituting from (30) into (29), and solving for $I_w{}^*$, the welfare maximizing total investment, we have

$$I_w{}^* = \max\left\{0, \frac{N}{\lambda\sigma_p{}^2[\beta(N-1)+1]}[(\bar{p}-r) - \lambda\sigma_{Mp}]\right\}. \tag{31}$$

Note that as β, the correlation among project returns, falls, the optimum social investment increases.

The inconsistency between "optimum" corporate investment and Pareto optimality

In general the aggregate investment resulting from the investment decisions made by value maximizing firms will be inconsistent with the optimum total investment and distribution of that investment as seen by all investors. Therefore the solution is non-Pareto optimum.

The general solution (22a-b) indicates that the investment will be equally divided among all firms having the option to take it only if the a_j's are identical for all of these firms. Thus, except in such a special case (and the case for which $\beta = 1$ in which investors are indifferent regarding the distribution), the distribution of the total investment will be nonoptimum since it will be unequally distributed among all potentially investing firms, and the greater is the diversity of the values of the a_j's, the greater will be the inequality of the distribution of investment.

In addition, if the project returns have a non-negative covariance with the non-project cash flows of the firms actually taking the investment, the total investment will be less than the welfare maximizing investment $I_w{}^*$. To see this let us sum (22b) over all investing firms (where z denotes the index set of investing firms whose total number is $n - k + 1$):

$$\begin{aligned} S = \sum_{j\in z} I_j{}^* &= \left(\frac{n-k+1}{2+(n-k)\beta}\right)\left(\frac{\bar{p}-r-\lambda\sigma_{Mp}}{\lambda\sigma_p{}^2}\right) \\ &\qquad - \frac{1}{\sigma_p{}^2[2+(n-k)\beta]}\sum_{j\in z}\sigma_{jp} \\ &\le \left(\frac{n-k+1}{2+(n-k)\beta}\right)\left(\frac{\bar{p}-r-\lambda\sigma_{Mp}}{\lambda\sigma_p{}^2}\right) \quad \text{if} \sum_{j\in z}\sigma_{jp} \ge 0 \\ &\le \left(\frac{n-k+1}{2+(n-k)\beta}\right)\left(\frac{\beta(N-1)+1}{N}\right) I_w{}^* \quad \text{using (31)} \\ &< I_w{}^*, \end{aligned} \tag{32}$$

since $(n - k + 1)/N < 1$ and

$$\frac{\beta(N-1)+1}{2+(n-k)\beta}$$

is less than unity for $\beta = 1$ and strictly increasing in β and is therefore strictly less than unity in the interval $0 \le \beta \le 1$. We shall examine several interesting special cases below, but first let us develop the social wealth maximizing criterion.

☐ **The social wealth maximizing criterion.** Given that the new investment project outlined above is available to some (or all) firms in the economy, our goal here is to determine the aggregate investment, I_s, and the distribution of that investment across all firms which maximizes $\Delta V_M - I_s$, the net increment in the total market value of all firms. Since all investors hold risky asset portfolios which differ only by a scale factor, this will maximize the net increment in wealth for each investor. This problem can be decomposed into two stages. In the first stage I_s is held constant and the problem is to determine the optimum distribution of the aggregate investment I_s among firms. In the second stage, the optimum value of I_s is determined.

The first-stage problem is to maximize[19]

$$\begin{aligned}\Delta V_M(I) - I_s = \frac{1}{r}[(\sum_j I_j)(\bar{\rho} - r) - \lambda(2I_j\sigma_{Mp} \\ + \beta\sigma_p^2 \sum_j \sum_{k \neq j} I_jI_k + \sigma_p^2 \sum_j I_j^2)]\end{aligned} \tag{33}$$

with respect to $\{I_j\}$, subject to: $\sum I_j = I_s$ and $I_j \geq 0, j = 1, 2, \ldots, N$.

The Lagrange equation for the first-stage problem is

$$\begin{aligned}L = \frac{1}{r}[(\sum_j I_j)(\bar{\rho} - r) - \lambda(2I_j\sigma_{Mp} + \beta\sigma_p^2 \sum_j \sum_{k \neq j} I_jI_k + \sigma_p^2 \sum_j I_j^2)] \\ + \gamma[I_s - \sum I_j],\end{aligned} \tag{35}$$

where γ is the Lagrangian multiplier. The maximum conditions are

$$\begin{aligned}\frac{\partial L}{\partial I_j} = \frac{1}{r}[(\bar{\rho} - r) - 2\lambda(\sigma_{Mp} + \beta\sigma_p^2 \sum_{k \neq j} I_k + \sigma_p^2 I_j)] - \gamma = 0, \\ j = 1, 2, \ldots, N,\end{aligned} \tag{36}$$

$$\frac{\partial L}{\partial \gamma} = I_s - \sum I_j = 0. \tag{37}$$

We can rewrite (36) as

$$\begin{aligned}\frac{1}{r}[(\bar{\rho} - r) - 2\lambda(\sigma_{Mp} + \beta\sigma_p^2 I_s + \sigma_p^2(1 - \beta)I_j)] - \gamma = 0, \\ j = 1, 2, \ldots, N.\end{aligned} \tag{38}$$

Solving (38) for I_j we find (for $\beta < 1$) that I_j is equal to a constant (call it Q) independent of j, and as before this implies $I_j = \frac{1}{N} I_s$, i.e., that the investment should be equally distributed across all firms.

---

[19] Equation (33) is obtained by summing (17) over all firms, j. Note in addition that the risk term in (33) has a simple interpretation as the total increment to the variance of the market cash flow $\tilde{D}_M$ caused by $\{I_j\}$:

$$\begin{aligned}\text{var}(\tilde{D}_M + \sum_j I_j\tilde{\rho}_j) - \text{var}(\tilde{D}_M) &= 2\,\text{cov}(\tilde{D}_M, \sum_j I_j\tilde{\rho}_j) + \text{var}(\sum_j I_j\tilde{\rho}_j) \\ &= 2 \sum_j \sum_k I_j\sigma_{kp} + \beta\sigma_p^2 \sum_{j \neq k} I_jI_k + \sigma_p^2 \sum_j I_j^2.\end{aligned} \tag{34}$$

Substituting this solution for $\{I_j\}$ into (33), using (37), and recognizing that

$$\sum_j \sum_{k \neq j} \left(\frac{1}{N^2}\right) I_s^2 = \left(\frac{N-1}{N}\right) I_s$$

and

$$\sum_j \frac{1}{N^2} I_s^2 = \frac{1}{N} I_s^2 ,$$

we can rewrite (33) solely in terms of I_s. The second-stage problem then is to maximize

$$\Delta V_M - I_s = \frac{1}{r}\left\{ I_s(\bar{\rho} - r) - \lambda \left[2I_s\sigma_{Mp} + \left(\frac{N-1}{N}\right) I_s^2 \beta \sigma_p^2 + \left(\frac{1}{N}\right) I_s^2 \sigma_p^2 \right] \right\} \tag{39}$$

with respect to I_s, subject to $I_s \geq 0$. Maximization of (39) yields the social wealth maximizing investment

$$I_s^* = \max \left\{ 0, \frac{N}{2\lambda\sigma_p^2[\beta(N-1)+1]}[(\bar{\rho} - r) - 2\lambda\sigma_{Mp}] \right\} . \tag{40}$$

Note that as the correlation among project returns falls, the optimum social investment increases. If $\beta = 1$, (40) reduces to the solution found earlier for investment by a single firm. If $\beta = 0$ the social wealth maximizing investment is N times the social optimum for the case in which $\beta = 1$.

In general the aggregate investment resulting from the decisions made by value maximizing firms will be inconsistent with social wealth maximization as well as inconsistent with social welfare maximization. Wealth maximization also requires uniform distribution of the investment across all firms and, as we saw before, this will occur only under special circumstances.

Examination of (40) and (18) indicates that if $\beta = 0$ and all the $\sigma_{jp} = 0$ (i.e., the project returns for a given firm are uncorrelated with the cash flows of all other firms), there is no inconsistency between the aggregate corporate investment and the social wealth maximizing investment; again a very special case.

However, if the project returns have on the average a positive covariance with the non-project cash flows of all firms,[20] there will be overinvestment in the project. That is, $\sum_{j=1}^{N} I_j^*$ will be greater than I_s^* and there will be overproduction of uncertainty. To see this, note that for a desirable project we can rewrite (40), the social wealth maximizing investment I_s^* as:

$$I_s^* = [\beta(N-1)+1]^{-1}\left[\sum_{j=1}^{N} a_j - \left(\frac{N-1}{2\sigma_p^2}\right)\sigma_{Mp} \right]$$

$$< [\beta(N-1)+1]^{-1}[\sum_j a_j] , \qquad \text{if } \sigma_{Mp} > 0$$

---

[20] A condition which is probably descriptive of most investment projects in an economy with business cycles.

$$= [\beta(N-1)+1]^{-1}\Big[\sum_{j\in z} a_j + \sum_{j\notin z} a_j\Big] \quad \text{(where } z \text{ denotes the index set of investing firms whose total number is } n-k+1\text{)}$$

$$\leq [\beta(N-1)+1]^{-1}\left\{\sum_{j\in z} a_j + \frac{\beta[(N-1)-(n-k)]\sum_{j\in z} a_j}{2+(n-k)\beta}\right\}$$

$$\left(\text{since for all } j \in z,\ a_j < \frac{\beta\sum_{j\in z} a_j}{2+(n-k)\beta}\right)$$

$$= \frac{[2+\beta(N-1)]\sum_{j\in z} a_j}{[\beta(N-1)+1][2+(n-k)\beta]}$$

$$\leq \frac{2[\beta(N-1)+1]\sum_{j\in z} a_j}{[\beta(N-1)+1][2+(n-k)\beta]} \tag{41}$$

$$= \sum_j I_j^* = S \qquad \text{from (22a) and (22b)},$$

where the limits of summation run from 1 to N unless specified otherwise.

Two special cases

Since the general solutions are somewhat complex, and therefore difficult to visualize, let us consider two interesting special cases. Consider a project for which $\beta = 1$ and $\sigma_{j_p} = 0$ for all firms (that is, all project returns are perfectly correlated between firms, and the project returns are uncorrelated with the pre-investment cash flows of all firms). If n firms ($n \leq N$) have the opportunity to take the project, each firm j will in equilibrium desire to invest an amount

$$I_j^* = \left(\frac{2}{n+1}\right)a = \left(\frac{2}{n+1}\right)\left(\frac{\bar{p}-r}{2\lambda\sigma_p^2}\right), \tag{42}$$

where a is given by (20). Hence the total investment by all firms in this project is

$$\sum_j I_j^* = 2\left(\frac{n}{n+1}\right)\left(\frac{\bar{p}-r}{2\lambda\sigma_p^2}\right). \tag{43}$$

However, (40) gives the social wealth maximizing investment in this case as

$$I_s^* = \left(\frac{\bar{p}-r}{2\lambda\sigma_p^2}\right), \tag{44}$$

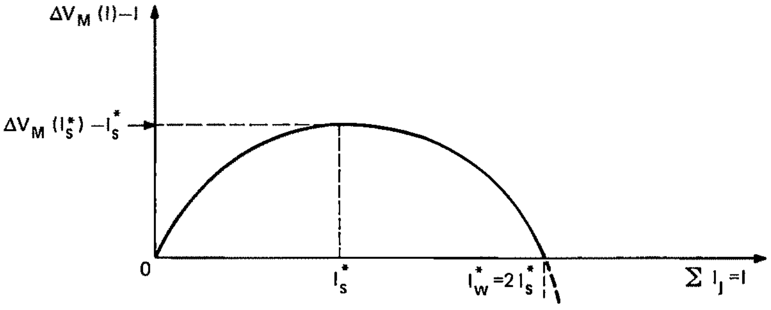
and by substitution into (43) we have

$$\sum_j I_j^* = 2\left(\frac{n}{n+1}\right)I_s^*. \tag{45}$$

That is, if more than a single firm have the opportunity to make this investment, the total investment made by all firms will be greater than that which will maximize the wealth of each investor. Indeed,

FIGURE 2

RELATIONSHIP BETWEEN THE NET CHANGE IN THE VALUE OF ALL ASSETS AND THE AGGREGATE CORPORATE INVESTMENT IN A PROJECT FOR WHICH $\beta = 1$ AND $\sigma_{jp} = \sigma_{kp} = 0$ FOR ALL j, k



we can see from (43) that if $n = 1$ the social wealth maximizing investment is achieved, and as n goes to infinity the total investment by all firms approaches twice the optimum, I_s^*. Furthermore, we can see by substitution into (39) that as n goes to infinity and $\sum_j I_j^*$ goes to $2I_s^*$ the net change in the market value of all assets, $\Delta V_M(I) - \sum_j I_j^*$, goes to zero, Note, however, that I_w^*, the welfare maximizing investment, for this special case is

$$I_w^* = \frac{\bar{p} - r}{\lambda \sigma_p^2} = 2I_s^*, \tag{46}$$

and as n goes to infinity the aggregate private investment goes to the social welfare maximizing investment. Figure 2 gives a graphical interpretation of this discussion. The aggregate investment by all firms will always lie between I_s^* and $I_w^* = 2I_s^*$ if $1 < n < \infty$.

There is one special case (other than the case of $n = 1$ for the project considered above) in which the maximizing behavior of firms will result in the maximization of social wealth. If we assume that the project returns are uncorrelated between firms as well as uncorrelated with the pre-investment cash flows of all firms (that is, $\beta = 0$ and $\sigma_{jp} = \sigma_{kp} = 0$) then substituting from (20) into (23) and summing over all j we have

$$\sum_j I_j^* = N\left(\frac{\bar{p} - r}{2\lambda \sigma_p^2}\right), \tag{47}$$

and from (40) we see that this is equal to the social wealth maximizing investment, I_s^*. However, since $I_s^* = \frac{1}{2} I_w^*$ we see that this solution is inconsistent with the attainment of maximum social welfare.

☐ **The question of optimality in the general case with large numbers of firms.** One can argue that the nonoptimality which we have discussed so far is due to the fact that our firms have monopoly power in the capital markets. Therefore it is of interest to ask what happens to the conclusion as we move to a competitive equilibrium. One way to do this is to examine the solution as we let the number of firms having access to the project increase (as we did above for the special cases where $\sigma_{jp} = 0$ and β equaled either one or zero). To do this let us consider holding the rest of the economy constant in size, i.e., fix $\tilde{D}_M$, and for simplicity let us assume that all the values of σ_{jp}, the covariance of the firms pre-project cash flows with the returns on the project, are equal. Now consider the limit of S, given in the first two lines of (32), as $(n - k + 1)$, the number of investing firms, goes to infinity. In the case of identical σ_{jp}, $n = N$, and we have

$$\lim_{N\to\infty} S = \frac{1}{\lambda\beta\sigma_p^{2}}[\bar{\rho} - r - \lambda\sigma_{Mp}] \tag{48}$$

for $0 < \beta \leq 1$, since if we are holding the size of the rest of the economy constant, $\sum_{j\epsilon z} \sigma_{jp}$ must be constant because $\sum_{j\epsilon z} \tilde{D}_j = \tilde{D}_M$, a constant with respect to N. (Note that our assumption of identical σ_{jp} implies $\tilde{D}_j = \dfrac{\tilde{D}_M}{N}$ and $\sigma_{jp} = \dfrac{\sigma_{Mp}}{N}$. However, this isn't strictly necessary for the results to hold.)

Now let us also examine the limit of the welfare maximizing solution for a desirable project, I_w^*, given by (31) as N goes to infinity:

$$\lim_{N\to\infty} I_w^* = \frac{1}{\lambda\beta\sigma_p^{2}}[\bar{\rho} - r - \lambda\sigma_{Mp}], \quad 0 < \beta \leq 1\,. \tag{49}$$

Thus, since (48) and (49) are identical, we see that as the number of firms approaches infinity the welfare maximizing level of investment goes to a limit and the aggregate investment made by value maximizing firms goes to this same limit. That is, if we were to compare two economies identical in every way except that the technology in one economy is such that the optimum size firm is much smaller (and therefore more firms exist), than in the other economy, the many-firm economy would be closer to the welfare maximizing investment I_w^* than the few-firm economy. These results seem consistent with the usual interpretation of the nature of a perfectly competitive equilibrium and monopolistic competition under certainty.

Note, however, that these results hold only for β, the between-firm correlation of project returns, in the range $0 < \beta \leq 1$. If $\beta = 0$, and all the $\sigma_{jp} = 0$, we have already shown above [equation (47)] that $\sum I_j^* = \frac{1}{2}I_w^*$, and it is easily seen that

$$\lim_{N\to\infty} \frac{S}{I_w^*} = \frac{1}{2}$$

for the case where $\beta = 0$ and the σ_{jp} are all identical. An intuitive economic explanation for the discontinuity seems to exist for the case of independent project returns, $\beta = 0$.

The two extreme cases seem easy enough to explain. When $\beta = 1$, the new assets being "produced" by different firms are literally identical. Thus we have a case of a perfectly homogeneous product, and it is not at all surprising that as the number of firms goes to infinity the socially optimum level of output is attained. Likewise, when $\beta = 0$, the new assets produced by one firm are completely differentiated from the new assets produced by another firm. This becomes particularly apparent if we refer back to equation (21), which expresses the desired investment of firm j as a function of the total investment of other firms. Note that if $\beta = 0$, $I_j{}^*$ is independent of the investment of other firms. Thus each firm is in a complete monopoly position in the sense that the decisions of other firms have no effect on the "profitability" of its investment. Increasing the number of firms in this case does not then tend to eliminate monopoly power.

The case of $0 < \beta < 1$ is somewhat puzzling since it involves a retention of some degree of "product differentiation," and hence one might expect, as in the case of $\beta = 0$, that letting the number of firms go to infinity would not eliminate nonoptimalities due to monopoly power. However, since S approaches a finite limit as N goes to infinity when $\beta > 0$, the investment level of the individual firm must approach zero. In price theoretic terms, this results from the fact that when $\beta > 0$, the "products" of other firms are in some degree substitutes for the "product" of a given firm. This, combined with the fact that the total "output" of "other" firms increases as N gets larger, forces the individual firm's "demand curve" $\left(\frac{\Delta V_j}{I_j}\right)$ down and to the left. In the limit, then, $I_j{}^* \rightarrow 0$ and marginal revenue becomes arbitrarily close to average revenue, a condition for maximum social welfare.

In order to provide additional intuitive understanding about the differences between the cases where $\beta > 0$ and $\beta = 0$ let us examine the behavior of the total social risk from the project as the number of firms varies. To do so we can represent the case for $0 < \beta < 1$ as a combination of the two extremes. Consider the multivariate normal case where we can express the per-dollar returns on each firm's project as a linear combination of a "market factor," $\tilde{\pi}$, and an error term, $\tilde{e}_j$, which is independent of $\tilde{\pi}$, independent across all firms, and has zero expectation:

$$\tilde{\rho}_j = a_j + b_j\tilde{\pi} + \tilde{e}_j.$$

Since in our example we have assumed β identical among all firms, we know a_j, b_j, and $\sigma^2(\tilde{e}_j)$ are identical for all j. Now it seems peculiar that for $b = 0$ we get a nonoptimum result but for any other value of b the aggregate investment of value maximizing firms leads to the welfare optimum as the number of firms goes to infinity. One way to see why this happens is to examine the behavior of the total social risk from the project, which, if we assume the $\sigma_{j\rho}$ are all zero, is just the variance of the total dollar returns from the project

$$\operatorname{Var}\left(\sum_j I_j{}^*\tilde{\rho}_j\right) = S^2b^2\sigma^2(\pi) + \frac{S^2}{N}\sigma^2(e) \tag{50}$$

(since if the $\sigma_{j\rho}$ are all zero we know $I_j{}^* = S/N$).

Now the limit of (50) as $N \rightarrow \infty$ is

$$\lim_{N\rightarrow\infty} \operatorname{Var}\left(\sum_j I_j{}^*\tilde{\rho}_j\right) = S^2(\infty)b^2\sigma^2(\pi)\,,$$

where $S(\infty)$ is the limit given by (48). The reason the variance contribution from the independent residual components $(S^2/N)\sigma^2(e)$

in (50) goes to zero in the cases for which $\beta > 0$ (and therefore $\beta \neq 0$) is that as the total investment S goes to the finite limit given by (48) the total investment by each firm goes to zero and the law of large numbers operates to eliminate the residual risk. (Equation (21) shows quite clearly that the investment of firm j decreases as the amount invested by other firms increases as long as $\beta > 0$.) Of course, in the case where $\beta = 1$ there is no residual risk and the total social risk increases monotonically to its limiting value given above. Therefore in the case where $0 < \beta < 1$ the law of large numbers operates to eliminate all risk due to the independent components e_j as $N \to \infty$; and as far as society (or the individual) is concerned it is almost as though these components didn't exist (the case where $\beta = 1$).

However, if $\beta = 0$, the aggregate corporate investment $S = \sum I_j^*$ increases linearly with N and so does the optimal social investment [cf. (47) and (31)], but the latter is always twice as large. Since the investment of each firm in this case is independent of the total level of investment of all other firms [see (21)] the total investment does not converge with increasing N and the law of large numbers cannot operate to eliminate any of the residual risk as it can in the case where $\beta > 0$. To see this we need only examine (50) recognizing that $\beta = 0$ implies $b = 0$ and using the solution for $S(N)$ given by (47) for the case where $\beta = 0$ and all the $\sigma_{jp} = 0$:

$$\operatorname{Var}\left(\sum_j I_j^* \bar{\rho}_j\right) = N\left(\frac{\bar{\rho} - r}{2\lambda\sigma_p^2}\right)^2 \sigma^2(e).$$

Thus in the case where $\beta = 0$, increasing the number of firms provides no additional diversification, since we also increase the total investment. These cases can be compared to the situation where (1) one takes a *given* dollar amount and spreads it across more and more independent securities while the total dollar variance goes to zero or (2) one invests one dollar in one security, two dollars in two securities, etc., and the total dollar variance does not converge to zero but increases in proportion to N, the number of securities.

4. Social and corporate investment criteria with θ assumed constant

■ **The social investment criteria.** We consider here the social investment criteria where the price of risk θ in (2), the ratio of mean excess returns to standard deviation, is assumed constant. As mentioned in Section 2, this is a case in which maximum social welfare is coincident with maximum social wealth. To see that this is true, consider the maximization of social wealth with respect to the allocation of investment in the new opportunity described in Sections 2 and 3. The post-investment market and firm cash flows are given, as before, by (14) and (15), and we wish to maximize

$$V_{M'} - I = \frac{1}{r}[\bar{D}_{M'} - Ir - \theta\sigma_{M'}] \qquad (51)$$

subject to $I_j \geq 0, j = 1, 2, \ldots, N$. The necessary conditions are

$$\bar{\rho} - r - \theta\frac{\sigma_{M'\rho(j)}}{\sigma_{M'}} \leq 0, \quad j = 1, 2, \ldots, N, \qquad (52)$$

where θ is assumed constant and

$$\sigma_{M'\rho(j)} = \sigma_{M\rho} + \sum_{k \neq j}^{N} I_k \beta \sigma_\rho^2 + I_j \sigma_\rho^2 . \tag{53}$$

Since $\lambda = \dfrac{\theta}{\sigma_{M'}}$, the condition for maximum social wealth given by (52) is identical to the condition for maximum social welfare given by (28). Thus the allocations implied by the two objectives are, in this case, identical.

Unfortunately (52) cannot be solved directly for $I_w^* = \sum_j I_j^*$, but nevertheless we shall be able to draw some conclusions from a comparison of these conditions with those obtained from the maximization of firm values. Note that the structure of (52) implies that if $I_w^* > 0$ the investment should be spread equally across all firms.

The optimum corporate investment is given by the maximization of $V_j - I_j$ subject to $I_j \geq 0$. The necessary conditions for a maximum are

$$\bar{\rho} - r - \theta \left[\frac{\sigma_{M'} \dfrac{\partial \sigma_{j'M'}}{\partial I_j} - \sigma_{j'M'} \dfrac{\partial \sigma_{M'}}{\partial I_j}}{\sigma_{M'}^2} \right] \leq 0 ,$$

which can be restated as

$$\bar{\rho} - r - \theta \left(\frac{\sigma_{M'\rho(j)}}{\sigma_{M'}} + \frac{\sigma_{M'}^2 \sigma_{j'\rho(j)} - \sigma_{j'M'} \sigma_{M'\rho(j)}}{\sigma_{M'}^3} \right) \leq 0 , \qquad j = 1, 2, \ldots, N , \tag{54}$$

where $\sigma_{j'\rho} = \sigma_{j\rho} + I_j \sigma_\rho^2$, and $\sigma_{j'M'}$ is given by (16). Again, (54) cannot be solved directly for the $\{I_j^*\}$ but we can readily see that to the extent that the second term inside the parentheses is not zero the social and corporate criteria need not agree. Since there is no reason why this term must in general be zero we have the possibility that the investment decisions made by value maximizing firms will not result in maximum consumer welfare.

Let us consider a project for which $I_w^* > 0$. Then the maximization of social welfare requires

$$I_j^* = \frac{1}{N} I_w^* , \qquad j = 1, 2, \ldots, N , \tag{55a}$$

$$\bar{\rho} - r - \theta \frac{\sigma_{M'\rho(j)}}{\sigma_{M'}} = 0 , \qquad j = 1, 2, \ldots, N . \tag{55b}$$

If in this case the "competitive" equilibrium, $I^* = (I_1^*, I_2^*, \ldots, I_N^*)$, corresponds to the welfare maximizing solution, i.e., $I_j^* = I_w^*/N$, $j = 1, 2, \ldots, N$, then from (55b) and the fact that for $I_j^* > 0$ the equality in (54) must hold,

$$\sigma_{M'}^2 \sigma_{j'\rho(j)} - \sigma_{j'M'} \sigma_{M'\rho(j)} = 0 \qquad \text{for all } j \text{ at } I^* \tag{56}$$

or

$$\sigma_{M'}^2 \left(\sigma_{j\rho} + \frac{1}{N} I_w^* \sigma_\rho^2 \right) - \left(\sigma_{jM} + I_w^* \sigma_{j\rho} + \frac{1}{N} I_w^* \sigma_{M'\rho} \right) \sigma_{M'\rho} = 0 , \qquad \text{for all } j . \tag{57}$$

But (57) cannot simultaneously hold for all j for any arbitrary specification of σ_{jp} and σ_{jM} for all j. Thus in general the investment decisions made by value maximizing firms will not be such that the welfare (and wealth) of all consumers is maximized.

5. Summary and conclusions

■ Given that the goal of the firm is to maximize its value net of investment we have derived the optimal corporate investment under certainty[21] within the context of a mean-variance general equilibrium model of the determination of the prices of capital assets. We have also derived the social welfare and social wealth maximizing investment criteria and demonstrated that the private criterion and both of the social criteria are in general inconsistent.

We also considered the simultaneous solution of the investment decisions of all firms regarding a prospective project and concluded that in this situation as well, the collective investment behavior of value maximizing firms leads to a level of total investment which is non-Pareto optimum for all cases in which the project returns are positively correlated with the sum of the pre-investment cash flows of all firms. In the case where the price of risk measured by the ratio of total excess mean returns to total variance is assumed constant the level of aggregate corporate investment will always be less than the social welfare maximizing investment and more than the social wealth maximizing investment.

We caution the reader to recall that most of our analysis has dealt with a capital market which is somewhat less than perfect. The individual firm in our model, to the extent that it has access to new investment projects whose returns are not perfectly correlated with the returns of projects available to other firms, enjoys some monopoly power in the sale of claims to its new projects. Even in the case of perfect correlation between the new project returns of different firms, monopoly power may still exist if the total number of firms to whom these projects are available is held at any finite level. This is equivalent to restricting entry into the business of exploiting these new opportunities. In the last part of Section 3, we investigated the effects of removing this entry restriction by letting the number of firms become infinite. It was then found that the private allocation of investment converges to the welfare optimum allocation in all cases except that in which there is complete independence among the new project returns of different firms. This result is surprising in some respects since it only requires that there be some positive amount of correlation (which may be far less than perfect correlation) between the new project returns of different firms for the private allocation of investment to converge to the optimum allocation.

References

1. FAMA, E. "Risk, Return, and Equilibrium: Some Clarifying Comments." *Journal of Finance*, Vol. 23, No. 1 (March 1968), pp. 29–40.
2. ———. "Risk, Return and Equilibrium." Report No. 6831, Center for Mathematical Studies in Business and Economics, University of Chicago, June 1968.

---

[21] For the case of constant stochastic returns to scale.

3. ———, and MILLER, M. "The Theory of Finance." New York: Holt, Rinehart & Winston, 1972.
4. FISHER, I. *The Theory of Interest.* New York: Augustus M. Kelley, 1930.
5. HIRSCHLEIFER, J. "On the Theory of Optimal Investment Decision." *Journal of Political Economy,* Vol. 66, No. 4 (August 1958), pp. 329–52.
6. LINTNER, J. "The Market Price of Risk, Size of Market and Investor's Risk Aversion." *The Review of Economics and Statistics,* Vol. 52, No. 1 (February 1970), pp. 87–99.
7. ———. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *The Review of Economics and Statistics,* Vol. 47, No. 1 (February 1965), pp. 13–37.
8. ———. "Security Prices, Risk, and Maximal Gains from Diversification." *Journal of Finance,* Vol. 20, No. 5 (December 1965), pp. 587–616.
9. LONG, J. B., JR. "Wealth, Welfare, and the Price of Risk." Forthcoming in *Journal of Finance,* Vol. 27, No. 2 (May 1972).
10. ———. "Consumption-Investment Decisions and Equilibrium in the Securities Market." Forthcoming in M. C. Jensen, ed., *Studies in the Theory of Capital Markets,* New York: Praeger Publishers, 1972.
11. MILLER, M. H. and MODIGLIANI, F. "Dividend Policy, Growth, and the Valuation of Shares." *Journal of Business,* Vol. 34, No. 4 (October 1961), pp. 411–33.
12. ———. "Some Estimates of the Cost of Capital to the Electric Utility Industry, 1954–1957." *The American Economic Review,* Vol. 56, No. 3 (June 1966), pp. 333–91.
13. ———. "The Cost of Capital, Corporation Finance and the Theory of Investment." *The American Economic Review,* Vol. 48, No. 3 (June 1958), pp. 261–97.
14. MOSSIN, J. "Equilibrium in a Capital Asset Market." *Econometrica,* Vol. 34, No. 4 (October 1966), pp. 768–83.
15. ———. "Security Pricing and Investment Criteria in Competitive Markets." *The American Economic Review,* Vol. 59, No. 5 (December 1969), pp. 749–56.
16. SHARPE, W. F. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance,* Vol. 19, No. 4 (September 1964), pp. 425–42.
17. VAN HORNE, J. C. *Financial Management and Policy.* Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1968.
18. WESTON, J. F. and BRIGHAM, E. F. *Managerial Finance.* 3rd ed. New York: Holt, Reinhart & Winston, 1969.

Telephone communications in Canada: demand, production, and investment decisions

A. Rodney Dobell, Lester D. Taylor, Leonard Waverman,
Tsuang-Hua Liu, and Michael D. G. Copeland

Research Associates
Institute for Policy Analysis
University of Toronto

This paper presents a portion of an econometric model of the telecommunications industry in Canada, specifically dealing with demand, production, and investment decisions in the telephone sector. Following an introductory sketch of the structure of the sector is an analysis of demand for service, broken down where possible into business and residence, local and toll, categories. Estimation follows the procedures established by Houthakker and Taylor; the results suggest the presence of significant elements of habit formation with high income elasticities in all classes of demand for service, but indicate a substantial degree of price elasticity only in household demand for toll service.

Analysis of production conditions includes estimation of production, input requirements, and productivity functions for one dominant firm (Bell Canada) and cross section estimates for the several members of the Trans-Canada Telephone System together. A Cobb-Douglas production function is found to provide a satisfactory fit to the statistical aggregates representing output and factor inputs, when a suitable proxy for technical change is included.

In the analysis of capital outlays, a straightforward lagged accelerator model is selected and yields a plausible pattern of response to changes in output levels.

The relationships estimated in this paper thus provide a consistent structure for that portion of an econometric industry model which links demand back to the factor inputs necessary to meet that demand and thence to the employment and investment decisions necessary to provide those inputs.

1. Introduction

■ Early in 1969, the Department of Communications of the Canadian government requested the Institute for Policy Analysis at the University of Toronto to contribute a statistical background study to be included as part of an extensive investigation by the so-called Telecommission. The Institute submitted its report, including statisti-

---

A. Rodney Dobell received B.A. and M.A. degrees from the University of British Columbia and the Ph.D. from M.I.T. Professor of Economics at the University of Toronto, he is presently on leave in Ottawa to create and direct a program in quantitative analysis for Canadian government employees. Dr. Dobell is coauthor of a work on mathematical theories of economic growth.

Lester D. Taylor, Associate Professor of Economics at the University of Michigan, received the B.A. degree from the University of Iowa and the Ph.D. from Harvard, where he also served as Assistant Professor. He has served with cal background on many aspects of telecommunications in Canada, in the summer of 1970.[1] This paper provides a condensed version of results relevant to the telephone industry, specifically on demand and supply conditions or investment decisions within that industry.

The paper is divided into six sections. Following this introductory section is a brief historical sketch of the industry. The next three sections analyze this historical record selectively for the determinants of demand for service, the conditions of production, and the nature of capital expenditure decisions. The analysis is conducted generally at an aggregate level, though separate relationships are estimated for the dominant firm (Bell Canada) in each case. The main points of the study are summarized individually within each section.

2. The setting

■ **The network.** This study deals with an industry providing a specific set of (marketed) services, namely telephone service (transmission of voice messages, both local and toll) and related services such as private wire service (dedicated lines from one point to another for voice, data, or remote program transmission by a specific user) or switched subscriber service (services using common trunking facilities and involving the switching of messages between selected points).

Local and long-distance telephone service throughout Canada is provided by more than 2,000 companies, both investor-owned and public, serving over 8 million telephones. The largest company is Bell Canada, which services almost all of the communities in the provinces of Ontario and Quebec. Each of the other provinces has a different primary system for local telephone service. Telephone service in the Yukon and the Northwest Territories, parts of Newfoundland, and northern sections of British Columbia is provided by Canadian National Telecommunications. By virtue of the heavy investment required to establish the distribution network, each company effectively holds a monopoly within its own territory and is subject to federal, provincial, or municipal regulation, depending upon the company's charter and circumstances.

Trunking of long-distance calls is provided by the Trans-Canada Telephone System (TCTS), which is an informal association consisting of eight full-member telephone companies (Alberta Government Telephones, the Bell Telephone Company of Canada, the British Columbia Telephone Company, Manitoba Telephone System, the Newfoundland Telephone Company, Ltd., the New Brunswick Telephone Company, Ltd., Maritime Telephone and Telegraph, and Saskatchewan Government Telephones). The Canadian Overseas Telecommunications Corporation (COTC), a Crown corporation

---

the Harvard Development Advisory as advisor to the Colombian government and is coauthor of a widely-read monograph on consumer demand in the U. S.

Leonard Waverman, Assistant Professor of Economics at the University of Toronto, received the B.A. degree from the University of Toronto and his doctorate from M.I.T.

Tsuang-Hua Liu, Assistant Professor at Eastern Washington College, studied in Taiwan and Alberta before earning the Ph.D. from the University of Toronto.

Michael D. G. Copeland, Instructor in economics at York University, Canada, received the B.A. in applied mathematics from the University of Toronto and did graduate work at M.I.T.

[1] See [15].

providing overseas connections, is an associate member. Every telephone system in Canada has access to this national and international network through the facilities of one of the member companies.

There are extensive transcontinental microwave systems in operation in Canada, placing Canada as the second highest among the world's users of microwave communications systems on a miles-per-capita basis. TCTS and Canadian National/Canadian Pacific (CN/CP) Telecommunications both provide transcontinental microwave systems for the transmission of telephone, television, data, and other communications. Other systems link British Columbia, Alberta, Ontario, and Quebec with the Far North. Some organizations such as provincial power authorities maintain private microwave systems for their own communications purposes. The two main Canadian television interests—the CBS and the CTV television networks—lease microwave facilities for the relay of television programs from coast to coast, and the Department of Communications has opened a microwave band for the use of various Canadian educational authorities in an instructional television system.

☐ **The ownership structure of the telecommunications industry.** Canada's telecommunications common carriers are largely Canadian-owned. For example, American Telephone and Telegraph Company (A.T.&T.) holds only approximately 2 percent of the equity in Bell Canada, the remainder being widely held by Canadian residents. However, Anglo-Canadian, a subsidiary of General Telephone and Electronics Corporation (GT&E) in the U. S., has majority control of British Columbia Telephone Company, and Quebec Telephone is also associated with GT&E. Of the TCTS members, Bell Canada owns approximately 50 percent of the New Brunswick Telephone Company, Ltd., 99.6 percent of Newfoundland Telephone Company, Ltd., and 52 percent of Maritime Telephone and Telegraph (but only 5 percent of its voting shares). Maritime Telephone and Telegraph in turn controls Island Telephone Company (a telephone company serving Prince Edward Island, but not a TCTS member).

The systems of the cities of Edmonton and Thunder Bay are municipally owned. The Manitoba Telephone System, Saskatchewan Government Telephones, and Alberta Government Telephones are provincially owned, while the systems in Nova Scotia, New Brunswick, and Prince Edward Island are privately owned with provincial charters. British Columbia Telephone Company and Bell Canada are investor-owned with federal charters.

Bell Canada has an agreement with A.T.&T. whereby for an annual fee the latter company provides advice and assistance on technical and operating matters. Nine major telephone organizations have similar agreements with Bell; British Columbia Telephone Company has a corresponding agreement with GT&E.

With respect to equipment manufacturers and suppliers, Bell Canada's wholly owned subsidiary, Northern Electric, has an agreement with Bell Canada requiring Northern to manufacture materials upon Bell's request, and, like Bell Canada and A.T.&T., Northern Electric and Western Electric (an A.T.&T. subsidiary) have a reciprocal agreement regarding information relating to the development and manufacture of telephone equipment. Likewise, British

Columbia Telephone Company has an ownership link with a manufacturing enterprise (Automatic Electric). Canadian Telephone and Supplies, Ltd. and Dominion Directory Company, Ltd. are two wholly-owned subsidiaries of Anglo-Canadian Telephone Company. General Telephone and Electronics International, Inc. (a subsidiary of GT&E) owns all of the outstanding shares in Sylvania Electric (Canada), Ltd., and the latter company in turn owns all of the outstanding shares of Lenkurt Electric Company of Canada, Ltd., an equipment supplier. Thus, although the major common carriers are Canadian, many of the peripheral operations appear to fall within large U. S. systems.

☐ **The regulatory bodies.** The growing importance of telecommunications is indicated by the creation, through the 1969 Government Organization Act, of a federal Department of Communications. The duties of the Minister of Communications involve matters relating to telecommunications and the development and utilization generally of communications undertakings, facilities, systems, and services for Canada.

Telephone and telegraph companies incorporated under the federal Parliament are subject to the jurisdiction of the Canadian Transport Commission in the matter of rates and practices under the provisions of the Railway Act, while the other major companies are subject to provincial regulation. The International Telecommunication Convention and regional agreements set down the procedures for international telegraph and telephone communications. Radio communications in Canada are regulated mainly under the Radio Act, the Canada Shipping Act, and international conventions on shipping, aviation, and telecommunications, and by regional agreements. All broadcasting for the public in Canada is regulated under the Broadcasting Act, which is administered by the Canadian Radio-Television Commission (CRTC).

CN Telecommunications is a federal Crown corporation, while CP Telecommunications is investor-owned, but both are federally regulated by the Canadian Transport Commission (CTC) under the Railway Act.

Bell Canada and British Columbia Telephone Company also operate under federal jurisdiction exercised by the CTC under the Railway Act and the Acts of Incorporation of these two companies. The other members of the TCTS (with the exception of Saskatchewan Government Telephones) operate under provincial regulation, exercised by means of an administrative board or commission whose powers generally flow from the Public Utility Act of that province. Saskatchewan Government Telephones reports directly to the government of Saskatchewan. The municipal systems, such as the City of Edmonton Telephone System, are generally subject to municipal regulation.

☐ **The regulation of rates.** The general principle governing the regulation of rates in the telecommunications industry is that they be "just and reasonable" and not be discriminatory against any person or company. Since there is no encompassing statutory test of justness and reasonableness, varying tests have been used by the regulatory bodies at different times and places. In the setting of rates and charges, it has been the practice of the particular regulatory board or commission to allow a certain rate of return on the "value" or rate base of the telecommunications firm in question. As examples, the following are descriptions of the various ways the rates of telephone companies across Canada have been determined at one time or another.

In the provinces of Newfoundland, Nova Scotia, Alberta, and British Columbia, the rates are tested by formulas using the rate of return on a rate base calculated on a valuation of plant in service plus allowances for items such as working capital. There is no specified test in New Brunswick, Quebec, and Manitoba, but the rates must be "just and reasonable." In Manitoba, there is a list of specific factors to be taken into consideration. In Prince Edward Island, the rate of return is calculated on a rate base fixed by the regulatory authority. In Saskatchewan, the test of rates for the rural telephone systems is that they be sufficient to pay operating and other costs. The municipal systems in Ontario must set rates sufficient to meet the payments of principal and interest on debentures. The larger systems are permitted to earn a prescribed rate of return on capital invested.

However, there is now some indication that regulation will not continue to be based on a permissive rate of return (however determined) as the sole test of reasonable rates. In a 1969 judgment regarding a Bell Canada rate increase application, the Canadian Transport Commission stated that the effects of continuing changes in the economy upon the present financial position of Bell

> . . . illustrate the fallacy of attempting to establish as a sole test of the justness and reasonableness of rates, a maximum permissive rate of return. . . . No statutory requirement exists for the fixing of a rate base and a rate of return on that base. Therefore our decision is not to fix any permissive level at this time, but we propose to maintain constant surveillance over Bell's affairs and take any steps that may in the future call for further relief or for remedial action.[2]

It may be noted here that the Railway Act (which governs the rates for CN Telecommunications, CP Telecommunications, Bell Telephone Company of Canada, and British Columbia Telephone Company, has recently been amended by Bill C-11, entitled "An Act to Amend the Railway Act," so that the CTC now possesses the authority to regulate tolls for use of previously unregulated services and facilities. Previously, "private wire" services—that is, the use of telegraph or telephone wires where no toll was charged to the public—were exempt from CTC regulation.

3. The overall model

■ The relationships developed in the present study were designed to form part of an econometric model of the telecommunications sector and reflect a particular conceptual scheme. Briefly, this scheme takes demand to be dependent on the general macroeconomic environment as reflected in population growth and movement, the level of national income, and the general level of prices, and also upon the prices of telecommunications services. (The latter consideration is sometimes felt to be of little importance in determining demand, but—as will be seen in the sequel—our results suggest otherwise.) Isolation of the determinants of demand and its components forms the subject of the next section.

---

[2] Bell Canada Application, Railway Transport Committee, Canadian Transport Commission, File C995.178, 1969.

We have assumed that the output observed is in fact equal to demand, which is tantamount to assuming that scarce capacity has not rationed supply over the period we study. Taking demand, and hence output, as given, the problem is to determine the amounts of inputs—labor, capital, and intermediate goods—required to produce this output. To do this, we have estimated (in a later section) conventional production functions, with allowances in one form or another for technological progress. In fact, however, stocks of durable capital goods must be assumed fixed in the short run, and we therefore take the production function to determine required levels of labor input.

Current output and demand forecasts establish expected future capital stock requirements, and these in turn determine the levels of investment outlays necessary to ensure that additional capacity be in place when it is needed. Our study of the patterns of investment expenditure is contained in the last section.

The remainder of the econometric model requires that one go on to investigate the implied costs of operation on the one hand and financial capital requirements on the other. Together, these considerations, through the regulatory process, determine rates to be charged for services, and hence condition estimates of future demand. In addition, the industry's expenditures for inputs and capital equipment can be traced back to suppliers, both domestic and foreign.

For present purposes, however, the focus is on the three conventional relationships central to any industry model: demand, production, and investment. The following sections analyze the implied links between the relevant variables.

4. Demand

■ The preceding sections provided an introduction to the telecommunications sector in Canada. This section and the two following deal with the historical record through an econometric analysis that attempts to relate the movements in the major economic variables describing or affecting demand conditions, production decisions, and investment outlays, respectively. This section outlines the analysis of the structure and determinants of demand and sets forth the main empirical results.

☐ **Models employed in the analysis.** The analysis is a straightforward application of the dynamic model of demand that Houthakker and Taylor (H-T) developed in their investigation of consumer demand in the United States.[3] This model has been the point of departure for all of the categories of revenue analyzed, and in nearly every instance it (or one of its variants) has given plausible results. Since the H-T model is by now fairly widely known, there seems little point in retracing its derivation.[4] This being the case, only a brief sketch of the model will be presented here, and this primarily to develop notation.

---

[3] See [33].

[4] *Ibid.*, pp. 9-29.

The structural form of the model (with time treated continuously) consists of the two equations

$$q(t) = \alpha + \beta s(t) + \gamma x(t) + \lambda p(t) \tag{1}$$

and

$$\dot{s}(t) = q(t) - \delta s(t), \tag{2}$$

where q, x, and p refer to consumption, income, and price, respectively, and s designates a state variable representing the depreciated residue (or "stocks") remaining from past consumption expenditures. Since this state variable is in general unobservable, it is eliminated through the use of equation (2), which connects the change in s to new expenditures and current depreciation. A translation to discrete time then yields the estimating form of the model

$$q_t = A_0 + A_1 q_{t-1} + A_2 \Delta x_t + A_3 x_{t-1} + A_4 \Delta p_t + A_5 p_{t-1}, \tag{3}$$

where $A_0, \ldots, A_5$ are functions of the structural coefficients α, β, γ, λ, and δ.[5]

In some instances, the alternative dynamic model formulated by Houthakker and Taylor has been found to yield the best results. Unlike the model just described, which can be interpreted as an attempt on the part of the consumer to bring his state variable into line with his income and price, the alternative model sees him doing this with regard to the flow. In continuous-time form, this model then consists of

$$\dot{q}(t) = \theta[\hat{q}(t) - q(t)] \tag{4}$$

$$\hat{q}(t) = \xi + \mu x(t) + \zeta p(t), \tag{5}$$

where q is the desired rate of consumption. The estimating form of the model is given by the equation[6]

$$q_t = A_0{}^* + A_1{}^* q_{t-1} + A_2{}^*(x_t + x_{t-1}) + A_3{}^*(p_t + p_{t-1}). \tag{6}$$

Houthakker and Taylor refer to this model as the "flow-adjustment" model, in contrast to the first which they call the "state-adjustment" model. The flow-adjustment model is suggested when the coefficients A_3 and A_5 in equation (3) are large relative to A_2 and A_4.

☐ **Data sources.** The major sources of data for the study have been *Telephone Statistics*, an annual publication of the Dominion Bureau of Statistics (DBS)[7] and Bell Canada. The latter very kindly made available to the study a four-way breakdown of Bell's revenues by local service and long distance, residence and business. To be able to estimate demand functions for each of these four categories taken separately obviously adds a great deal of interest to the analysis. A final source of data has been the unpublished paper of R. E. Olley,[8] which itself employs data provided by Bell Canada.

In principle, the dependent variable in the equations should be the number of message units, where each message unit is weighted by the product of its time duration and geographical distance covered. However, neither of these last two items of information is available. This being the case, there has been no alternative but to use price-deflated revenue as the measure of output. The price indexes employed for this have been obtained from two sources: the DBS and total Bell Canada series have been deflated by the deflators of Bell Canada revenues which are given in the Olley paper, while the local-service and long-distance revenue series of Bell have been deflated by price indexes constructed from historical rate schedules filed with the CTC and made available to us by Bell Canada. Descriptions of these indexes are given a little later on.

---

[5] For the relationships connecting the two sets of coefficients, see [33], equations (1.39) and (1.41)-(1.44).

[6] The coefficients $A_0{}^*, \ldots, A_3{}^*$ are connected to ξ, μ, ζ, and θ via equations (1.75)-(1.80) in [33].

[7] See [16].

[8] See [54].

☐ Empirical results for total Canada, DBS data.

1. Total revenue

The first category analyzed is the total revenue of all telephone companies in the country. This category refers to combined revenue from all sources, which means that revenues from directory advertising are included in addition to those from normal telecommunications activity.[9] Since DBS does not publish deflators for the revenue data that it collects, it has been necessary to deflate by an implicit deflator for the total revenue of Bell Canada.[10] However, given the importance of Bell's revenues in the total for the industry, the error introduced by this procedure *should* be of little consequence.[11] For the income variable, we have used GNP measured in constant (1967) dollars, and for the relative price we have used the Bell deflator divided by the implicit deflator for GNP. The data are annual and cover the period 1952–1967. Ordinary least-squares estimation has been employed, not only for this category but throughout the entire paper.

The equation estimated is as follows (*t*-ratios are in parentheses):

$$\begin{aligned} q_t = \underset{(1.28)}{303.21} + \underset{(6.54)}{0.9176}q_{t-1} + \underset{(0.70)}{0.00131}(x_t + x_{t-1}) \\ - \underset{(-\,1.98)}{1.3313}(p_t + p_{t-1})\,, \end{aligned} \tag{7}$$

where

q = total telephone company revenue (in millions)/implicit deflator for total revenue of Bell Canada,

x = GNP in millions of 1967 dollars,

p = implicit deflator for total revenue of Bell Canada/implicit deflator for GNP (1967 = 100), and

$R^2 = 0.999$, $S_e = 8.89$, $DW = 1.78$, $\hat{\theta} = 0.316$,

$\xi = 3679.7$, $\hat{\mu} = 0.032$, and $\hat{\zeta} = -\,32.31$.

---

[9] Strictly speaking, advertising revenues should be netted out of total revenue, but since DBS does not break out advertising revenues separately this cannot be done. The closest that one can come would be to use the sum of local-service and long-distance revenues, but this would purge too much. Excluded would be revenues generated by leased lines, and these should clearly be part of total revenue. For this reason, it was elected to use total revenue from all sources. Since directory advertising probably moves closely with local revenues, the error involved should be small.

[10] See [54].

[11] However, the results for Bell Canada, discussed below, suggest that this may not be the case.

The elasticities calculated at 1967 levels are:[12]

	Income	Relative Price
Short-run	0.56	-0.88
Long-run	1.77	-2.78 .

This is an instance of the flow-adjustment model. The items worthy of note in the equation are the substantial long-run income and price elasticities (both are considerably greater than unity in absolute value), the high R^2, and the absence of autocorrelation in the residuals (as evidenced by a value of 1.78 for the Durbin-Watson coefficient).[13] Despite the tight fit, however, the income and price coefficients are not as significant statistically as one would like. The t-value for income is less than one, while that for the relative price barely borders on 2. This equation was selected in preference to one based on the stock adjustment model because of an implausibly large long-run price elasticity with the latter; otherwise, the two models give essentially the same results.

Despite the fact that the income coefficient is insignificant by conventional statistical standards, the implied (long-run) income elasticity of 1.8 seems reasonable. However, since a substantial part of total revenue consists of revenues from local service—which we should expect to be relatively independent of price—a price elasticity of -2.8 may be too high. The reasons for our holding this view will become clear in a moment, but here we should note a technical reason as to why our procedure might overestimate the price elasticity. The constant-dollar revenue series is obtained by a deflator that forms the numerator of the relative price. If for some year (or years) this deflator is too low, our measure of constant-dollar revenue for that year will be too high, and the regression will tend to attribute this high expenditure to the (incorrectly) low price. And the same bias will operate on the other side if the deflator for some years is too high. Thus, the high price elasticity in equation (7) may reflect simply the way in which constant-dollar revenue has been derived.[14]

In addition to equation (7), an equation has been estimated with revenue and income expressed in per-capita terms, to see how much the aggregate income elasticity is influenced by population growth. This is important information, for if the rate of growth of the population should taper off, projections based on equation (7) could be seriously biased. Since equation (7) is estimated with aggregate revenue and aggregate income, the income elasticity reflects an amalgam of population growth and higher income per head. If the individual income elasticity were unity, then whether the variables were in aggregate or per-capita terms would not matter, but when the individual elasticity is greater than unity—as is the case with expenditures for telephone services—it does matter. If the goal is to estimate an elasticity applicable to a single individual, population growth will lead to an elasticity estimated from aggregate data which is biased toward unity.

---

[12] With the flow-adjustment model, the short-run elasticities are defined as θ times the long-run elasticities ($= \mu(q/x)$, etc.).

[13] It must be kept in mind, however, that with models having the lagged dependent variable as a predictor the Durbin-Watson coefficient is biased toward 2. See [52].

[14] Of course, only when errors in measurement of the price variable are large will this bias be significant. In the absence of such errors, no such bias is entailed.

The results from applying the stock-adjustment model to total revenue per capita follow:

$$\begin{array}{rcccccc} q_t = & 6.50 & + & 0.9232q_{t-1} + & 0.00339x_t & - & 0.0882p_t, \\ & (0.64) & & (19.34) & (1.77) & & (-1.62) \end{array} \tag{8}$$

where

q = total telephone revenue per capita/implicit deflator for total revenue of Bell Canada,

x = GNP per capita in 1967 dollars,

p = implicit deflator for total revenue of Bell Canada/implicit deflator for GNP (1967 = 100), and

$R^2 = 0.999, \quad S_e = 0.40, \quad DW = 1.76, \quad \hat{\alpha} = 3.38, \quad \hat{\beta} = 1.920,$

$\hat{\gamma} = 0.0018, \quad \hat{\lambda} = -0.46,$ and $\hat{\delta} = 2\,.$

The elasticities computed at 1967 levels are:

	Income	Relative Price
Short-run	0.10	−0.08
Long-run	2.48	−2.01 .

Since this is the first equation presented using the stock-adjustment model, it will be useful to discuss it in some detail. First, it should be noted that the equation is actually a special case of the stock-adjustment model, for it has been estimated under the restriction that coefficients A_2 and A_4 are equal, respectively, to coefficients A_3 and A_5. This form of the model implies a substantial element of habit formation, a fact evidenced in income and price elasticities which are large in the long run but nearly zero in the short run. This property reflects the strong inertia which is characteristic of the demand for services. The statistical qualities of this equation parallel those for the aggregate equation. The coefficients have the correct sign, the fit is good, and again there is no evidence of autocorrelation in the residuals. Income is somewhat more significant statistically and the relative price somewhat less so, but once again neither has a t-value greater than 2. The long-run income elasticity of 2.5 is well in excess of its value in the aggregate equation. However, the long-run price elasticity is lower than in the aggregate equation.

2. Local-service revenues

The second DBS telephone revenue category analyzed is revenues from local service. As with total revenue, the data refer to the combined revenue from local service of every telephone company in the country. The data have been deflated with the implicit local-service deflator for Bell Canada, and this variable also forms the numerator of the relative price. Once again, the time period covered is 1952–1967. The empirical results follow:

$$\begin{array}{rcccc} q_t = & 21.76 & + & 0.9129q_{t-1} + & 0.00088(x_t + x_{t-1}) \\ & (108.04) & & (10.19) & (1.23) \\ & & & & - \; 0.1876(p_t + p_{t-1}), \\ & & & & (-0.59) \end{array} \tag{9}$$

where

q = total local-service telephone revenue (in millions)/implicit deflator for local-service revenues of Bell Canada,

x = GNP in millions of 1967 dollars,

p = implicit deflator for local-service revenues of Bell Canada/implicit deflator for GNP (1967 = 100), and

$R^2 = 0.999$, $S_e = 2.98$. $DW = 1.87$, $\hat{\theta} = 0.333$, $\hat{\xi} = 249.8$, $\hat{\mu} = 0.020$, and $\hat{\zeta} = -4.31$.

The elasticities computed at 1967 levels are:

	Income	Relative Price
Short-run	0.72	−0.23
Long-run	2.16	−0.70 .

Since local service accounts for more than 50 percent of total telephone revenue, we should expect the local-service results not to stray too far from those for total revenue, and this is the case. The flow-adjustment model gives the best results, the fit is extraordinarily tight, the residuals appear free of autocorrelation, and the income elasticity, though of only marginal statistical importance, is large. There is a major difference with regard to the importance of the relative price, but this is hardly surprising. On the part of households, local-service revenues should be determined primarily by the number of families and the level of personal disposable income, while on the part of businesses the general level of economic activity should be the primary determining factor. In neither case should local-service rates be expected to play much of a role, and, indeed, local-service revenues are observed to be inelastic with respect to price.

3. Long-distance revenues

For long-distance revenues, the same form of the stock-adjustment model as was used for total revenue per capita—coefficient A_3 equal to A_2, and coefficient A_5 equal to A_4—has been found to yield the best results. The equation is as follows:

$$\begin{array}{l} q_t = \underset{(0.93)}{112.15} + \underset{(6.21)}{0.9200q_{t-1}} + \underset{(0.79)}{0.00104x_t} - \underset{(-1.36)}{0.9478p_t}, \end{array} \qquad (10)$$

where

q = total long-distance telephone revenue (in millions)/implicit deflator for long-distance revenues of Bell Canada,

x = GNP in millions of 1967 dollars,

p = implicit deflator for long-distance revenues of Bell Canada/implicit deflator for GNP (1967 = 100), and

$R^2 = 0.997$, $S_e = 6.82$, $DW = 1.91$, $\hat{\alpha} = 58.41$, $\hat{\beta} = 1.917$, $\hat{\gamma} = 0.0005$, $\lambda = -0.494$, and $\hat{\delta} = 2$.

The elasticities calculated at 1967 levels are:

	Income	Relative Price
Short-run	0.08	−0.11
Long-run	1.90	−2.57 .

Apart from income and relative price not being as significant statistically as one would prefer, this equation presents a plausible picture of the demand for long-distance service. Habit formation is strong, and, in contrast to local-service revenues, there is a sharp long-run response of demand to a change in price. The long-run elasticity with respect to income is also large. In keeping with the general results of this section, the equation fits the historical data extremely well, and once again there is no problem of autocorrelated residuals, at least so far as the DW statistic can be accepted.

☐ **Empirical results by region, DBS data.** In a country as geographically and economically diverse as Canada, there might be some question as to whether the demand functions estimated in the preceding section for the country as a whole are also applicable to the separate regions of the country in isolation. Since DBS provides a breakdown of its telephone revenue data by province, it is reasonably straightforward to test whether this is the case. To do this, we divided the country into four regions, namely:

(1) The Maritime Provinces,
(2) Ontario and Quebec,
(3) The Prairie Provinces, and
(4) British Columbia and the Yukon Territory.

For each, an equation such as (7) was estimated in per-capita terms, along with a "total" equation based on the data from all four regions pooled. The hypothesis that the demand for telephone service is homogeneous among regions may then be tested by an analysis of covariance. On the basis of an F-test (or a "Chow" test)[15] performed on the reduction in residual variance in going from the "total" equation to the groups of "individual" equations, we cannot reject this hypothesis at a 90-percent significance level.

☐ **Empirical results, Bell Canada data.** On the basis of data made available to the larger study by Bell Canada, it has been possible to extend the analysis considerably beyond that possible with DBS data alone. Bell provided a breakdown of its revenue between residence and business and a further breakdown within each of these categories between local service and long distance. This detail is especially useful because, in addition to the difference in demand between local service and long distance, there are almost certainly differences in the demand between each of these services for businesses and households. In particular, we should not be surprised to find that the substantial long-run elasticity for long distance is accounted for by households rather than by businesses.

Unfortunately, the Olley paper (from which we obtained the Bell Canada price deflators)[16] does not provide separate deflators for business and residence revenues. Rather than use the same deflator for both, we have constructed our own on the basis of historical rate schedules filed with the CTC and made available to us by Bell Canada. For local service, the residence price index is based on the monthly rate for a two-party line in cities of over 250,000, and the business price index is based on the monthly rate for an individual line, also in cities of over 250,000. For long distance, the residence price index is based on the nighttime station-to-station rate beginning at 6 P.M. for a 3-minute 350-mile call,[17] while the business price index is based on the daytime station-to-station rate for the same distance.

---

[15] See [10].

[16] See [54].

Since Bell Canada operates primarily in the most densely populated areas of Ontario and Quebec, the income variables should reflect the income level of these two provinces. For households, the income measure should be personal disposable income, which we have, but for businesses the measure ought to be GNP (or better still, gross business product). However, as noted above, DBS does not publish GNP by province and so we have had to use personal disposable income in the business equations also. With one exception the equations which follow are based on data covering the period 1950–1967; the exception is the equation for total revenue of Bell, which is based on data for 1952–1967. Finally, the residence equations are estimated in per-capita terms.

1. Residence local-service revenues

$$\begin{array}{llll} q_t = & -9.39 \;+ & 0.9494q_{t-1} + & 0.000424(x_t + x_{t-1}), \\ & (-0.77) & (20.28) & (1.47) \end{array} \qquad (11)$$

where

q = residence local-service revenues of Bell per capita/local-service price index for residences (1967 = 100),

x = personal disposable income per capita for Ontario and Quebec in 1967 dollars, and

$R^2 = 0.999$, $S_e = 0.10$, $DW = 1.68$, $\hat{\theta} = 0.197$, $\hat{\xi} = -7.71$, and $\hat{\mu} = 0.017$.

The income elasticities computed at 1967 levels are:

Short-run	0.47
Long-run	2.38 .

The important finding for this category is the absence of relative price as an explanatory factor. When price was included in the model, its sign was correct, but t-values on all coefficients in the equation, with the exception of the one for q_{t-1}, were less than one, and when price was tried by itself, its sign became positive (though not significantly so). Consequently, the relative price was excluded and income retained instead.

2. Residence long-distance revenues[18]

$$\begin{array}{lllll} q_t = & 0.24 \;+ & 0.8674q_{t-1} + & 0.00117\Delta x_t + & 0.00103x_{t-1} \\ & (0.14) & (2.77) & (1.02) & (0.65) \\ & & & -\;0.0248\Delta p_t - & 0.00927p_{t-1}, \\ & & & (-4.04) & (-1.01) \end{array} \qquad (12)$$

---

[17] To avoid difficulties with the introduction of different rate classes beginning at different times through the night, we have computed the rate for a call placed at 7 P.M.

[18] This equation was estimated without the benefit of a program which takes into account the identifying restriction, $A_2A_6 = A_3A_4$, on δ when both income and price are predictors and the value of δ is not set *a priori*. This means, therefore, that (12) in fact yields two estimates of δ, one from $\hat{A}_2$ and $\hat{A}_3$ and the other from $\hat{A}_4$ and $\hat{A}_5$. The value of δ reported is a weighted average of these two estimates, the weights being the inverses of the t-ratios of $\hat{A}_3$ and $\hat{A}_5$. For a discussion of this identification problem, see pp. 47–51 in [33].

where

q = residence long-distance revenues of Bell per-capita/long-distance price index for residences (1967 = 100),

x = personal disposable income per capita for Ontario and Quebec in 1967 dollars,

p = long-distance price index for residences/implicit deflator for personal consumption expenditure (1967 = 100), and

$R^2 = 0.993$, $S_e = 0.15$, $DW = 1.95$, $\hat{\alpha} = 0.282$, $\hat{\beta} = 0.7579$, $\hat{\gamma} = 0.0007$, $\hat{\gamma}' = 0.0044$, $\hat{\lambda} = -0.149$, $\hat{\lambda}' = -0.1367$, and $\hat{\delta} = 0.8996$.

The elasticities calculated at 1967 levels are:

	Income	Relative Price
Short-run	0.20	−0.30
Long-run	1.27	−1.90.

In contrast to local service, the household demand for long-distance service is seen to be quite sensitive to price, for not only is the relative price an important predictor statistically, but its (long-run) elasticity is well in excess of 1 (in absolute value). The response of long-distance demand to a change in income is also elastic in the long run, but not as decidedly so as the demand for local service. Apart from the barely significant income coefficient, the statistical aspects of this equation are quite good. The fit is tight, and the residuals appear to be free of autocorrelation.

3. *Total residence revenue*

$$\begin{array}{ccccc} q_t = & 2.96 + & 0.8641q_{t-1} + & 0.00232x_t - & 0.353p_t, \\ & (0.74) & (7.95) & (1.10) & (-1.87) \end{array} \qquad (13)$$

where

q = total residence revenue of Bell per capita/price index for residential telephone service (1967 = 100),[19]

x = personal disposable income per capita for Ontario and Quebec in 1967 dollars,

p = price index for residential telephone service/implicit deflator for personal consumption expenditure (1967 = 100), and

$R^2 = 0.996$, $S_e = 0.27$, $DW = 2.13$, $\hat{\alpha} = 0.159$, $\hat{\beta} = 1.854$, $\hat{\gamma} = 0.0012$, $\hat{\gamma}' = 0.0044$, $\hat{\lambda} = -0.189$, $\hat{\lambda}' = -0.1367$, and $\delta \doteq 2$.

The elasticities calculated at 1967 levels are:

	Income	Relative Price
Short-run	0.12	−0.09
Long-run	1.63	−1.19.

We should expect the results for total residence revenue to lie intermediate between those for local service and those for long distance, as observed. Perhaps the most relevant finding with this equation is that the price effect for long-distance calls is apparently of sufficient strength to make the household demand for telephone services, local and long-distance combined, slightly elastic (in the long run) with respect to price.

---

[19] Constructed as $0.675\ pl_s + 0.325\ pl_d$, where ls and ld refer to local service and long distance, respectively, and where the weights are an average of the initial and final shares of each revenue category in total residence revenue.

4. Business local-service revenues

$$\underset{}{q_t} = \underset{(4.09)}{-\,29.72} + \underset{(6.64)}{0.6787q_{t-1}} + \underset{(4.00)}{0.00236(x_t + x_{t-1})}, \qquad (14)$$

where

q = business local-service revenues of Bell Canada (in millions)/ price index for business local service,

x = personal disposable income for Ontario and Quebec in millions of 1967 dollars, and

$R^2 = 0.999$, $S_e = 2.35$, $DW = 1.50$, $\hat{\theta} = 0.383$, $\hat{\xi} = -\,92.50$, and $\hat{\mu} = 0.015$.

The income elasticities calculated at 1967 levels are:

Short-run	0.59
Long-run	1.55 .

We see that, as with residences, business demand for local service is indicated to be independent of price. This is, of course, a sensible result and one which telephone companies appear to have taken into account in setting local-service rates for businesses vis-à-vis those for households. Despite a hint of positive autocorrelation in the residuals (as indicated by the Durbin-Watson coefficient of 1.50), the statistical quality of this equation is very good. All coefficients are multiples of their standard errors, and, as usual with these data, the R^2 is very high.

5. Business long-distance revenues

$$q_t = \underset{(-1.47)}{-\,22.75} + \underset{(5.74)}{0.8708q_{t-1}} + \underset{(1.59)}{0.00213x_t}, \qquad (15)$$

where

q = business long-distance revenues of Bell Canada (in millions)/ price index for business long-distance calls,

x = personal disposable income for Ontario and Quebec in millions of 1967 dollars, and

$R^2 = 0.990$, $S_e = 5.01$, $DW = 2.37$, $\hat{\alpha} = -\,12.16$, $\hat{\beta} = 1.862$, $\hat{\gamma} = 0.0011$, $\hat{\gamma}' = 0.016$, and $\delta = 2$.

The income elasticities calculated at 1967 levels are:

Short-run	0.19
Long-run	2.76 .

At first glance, this equation might seem to present a rather surprising result, for once again the relative price is absent as a predictor. However, upon reflection, we should not find this so surprising. For many businesses, the long-distance call has become an indispensable instrument of commerce. Even if long-distance rates were several times their present levels the phone might still, in many instances, be the least costly mode of communication, especially when the time element is taken into account. Lacking relative price indexes to measure costs of relevant substitutes for such communications activities, we probably miss some of the price sensitivity in these business decisions and therefore perhaps overestimate the long-run income elasticity.

6. Total business revenue

$$\begin{array}{lcccc} q_t = 185.53 & + \; 0.5433q_{t-1} & + \; 0.00652x_t & - \; 1.5370p_t, \\ \quad (2.80) & (3.71) & (2.88) & (-3.45) \end{array} \qquad (16)$$

where

q = total business revenue of Bell Canada (in millions)/price index for telephone service for businesses,[20]

x = personal disposable income for Ontario and Quebec in millions of 1967 dollars,

p = price index for business telephone service/implicit deflator for GNP (1967 = 100), and

$R^2 = 0.998,\quad S_e = 4.87,\quad DW = 1.67,\quad \hat{\alpha} = 120.22,\quad \hat{\beta} = 1.408,$

$\hat{\gamma} = 0.0042, \hat{\gamma}' = 0.014, \hat{\lambda} = -0.996, \hat{\lambda}' = -3.365$, and $\delta = 2$.

The elasticities calculated at 1967 levels are:

	Income	Relative Price
Short-run	0.27	−0.39
Long-run	0.93	−1.33 .

This equation presents a number of anomalies. Statistically, it it perhaps the best equation in the paper, for all coefficients are at least twice their standard errors, the R^2 is very high, and there is only a hint of autocorrelation in the residuals. The problem, however, is that the equation is seriously out of line with the equations for local service and long distance; the relative price is absent from both equations (14) and (15) but shows up strongly in equation (16). And, probably not independent of this, the (long-run) income elasticity for the combined revenues is less than for either of the components. While it is quite possible to explain this as a statistical phenomenon, an economic interpretation is another matter. Thus, we must leave it as a statistical quirk.

7. Total revenue of Bell Canada

The following equation was estimated to describe total Bell Canada revenue:

$$\begin{array}{llll} q_t = 18.52 & + \; 0.7594q_{t-1} & + \; 0.00492(x_t + x_{t-1}) & \\ \quad (0.15) & (4.64) & (1.62) & \\ & & & - \; 0.3266(p_t + p_{t-1}), \\ & & & (-0.94) \end{array} \qquad (17)$$

where

q = total revenue of Bell Canada (in millions)/implicit deflator for Bell total revenue (from Olley study),

x = personal disposable income for Ontario and Quebec in millions of 1967 dollars,

---

[20] Calculated as $0.617pl_s + 0.383pl_d$, where ls and ld refer to local service and long distance, respectively, and where the weights again are averages of the initial and final shares of the revenue categories concerned in total business revenue.

p = implicit deflator for Bell total revenue/implicit deflator for GNP (1967 = 100), and

$R^2 = 0.999, \quad S_e = 4.92 \quad DW = 2.46, \quad \hat{\theta} = 0.274, \quad \hat{\xi} = 76.97,$

$\hat{\mu} = 0.041,$ and $\hat{\zeta} = -2.71.$

The elasticities calculated at 1967 values are:

	Income	Relative Price
Short-run	0.43	−0.11
Long-run	1.57	−0.39 .

We have presented the equation for the total revenue of Bell Canada last rather than first in order to illustrate better why it is both inaccurate and misleading to speak of a single all-encompassing demand for telephone services. The preceding subsections have demonstrated quite clearly that there are at least four unambiguously defined submarkets for telephone communications, and we would probably find even more if further disaggregation of the data were possible. The demand for telephone services for all customer and service categories combined thus reflects the interplay of each of the submarkets studied, and this must be kept in mind in interpreting equation (17), and more particularly in assessing the equation for total revenue estimated from DBS data [equation (7)].

Perhaps the most significant aspect of equation (17) is the weakness, both statistically and quantitatively, of the relative price. While this is in keeping with other equations for Bell Canada—all, that is, except for equation (12)—it is quite at variance with equation (7). Since Bell revenues are a substantial part of the total, one would expect (especially in view of the results of Section 2, which suggest that the demand for telephone services in Canada is homogeneous with respect to region) that equations (7) and (17) would be in closer agreement. The income elasticities agree quite well, but the (long-run) price elasticity in equation (7), it will be recalled, is well in excess of 2. Since, strictly speaking, the Bell implicit deflator is appropriate only to that part of total revenue which is Bell's, the problem probably lies in equation (7) rather than in equation (17). This is because, as was noted at the time that equation (7) was being discussed, incorrect deflation will lead to an exaggerated estimate of the price elasticity. Since this is the case, we should probably put more confidence in the price elasticity from equation (17) than in the one from equation (7).

☐ **Summary of results.** In conclusion, it is useful to summarize the major results of this section:

(1) The demand for telephone services is characterized by strong habit formation. This is reflected in a response in demand to a change in income or relative price that is small in the short run compared with the long run.

(2) The long-run elasticity of demand in Canada with respect to income is well in excess of unity. This is true not only in the aggregate but also for residences (both local service and long distance) and businesses (both local service and long distance) taken separately.

(3) The demand for local service for both households and businesses and long-distance service for businesses appears to be independent of relative price.

(4) On the other hand, the demand for long-distance services on the part of households displays a substantial long-run price elasticity.

(5) The demand for telephone services appears to be homogeneous across regions.

(6) The statistical quality of the estimated equations varies from excellent to only fair. The fits of the equations are extraordinarily tight, even for time-series data, and there are no apparent problems of autocorrelated residuals. The coefficients of income and price are often not as significant statistically as one would prefer, but much of the time this can be attributed to strong trends in the data which introduce an undesirable amount of intercorrelation into the explanatory variables.

5. Production

■ This section attempts to estimate the contributions of each of the factors of production to the output of the telephone industry. In attempting to define the productive characteristics of the industry we have estimated three types of relationship: a production function showing output as a function of inputs, input requirements functions displaying a specific input as a function of output, and productivity functions giving inputs as functions of relative factor prices.

For the production function, the form yielding the best estimates proved to be the "Cobb-Douglas" production function,[21]

$$Y_t = AL_t^{\alpha}K_t^{\beta}D_t, \tag{18}$$

where Y, L, and K are measures of output, labor services, and capital services, respectively, while D_t represents the technology at time t. The two parameters of primary interest, the elasticity of substitution, σ, and the returns-to-scale parameter, γ, are trivial to estimate in this form, since use of the Cobb-Douglas production function implies a constant elasticity of substitution equal to one, and the returns-to-scale parameter γ is given by $\gamma = \alpha + \beta$.

By many standards, the results obtained by use of this estimation procedure were adequate. Nevertheless, certain inconsistencies between results for various industry subgroups were troubling, and the use of input requirements functions suggested itself as an alternative.

While a production function estimates the maximum output attainable with given inputs, one can also summarize the technology of the production process through a set of functions which indicate, for a given amount of output, the minimum inputs required to produce that output. These are referred to as input requirements functions and might be particularly useful in an industry like telecommunications where the level of output is often thought to be primarily determined by factors external to the industry.

---

[21] It is interesting that the simple Cobb-Douglas form yielded good results. Investment in the industry is lumpy, and some types of equipment have long lead times between order and delivery. Capital is long-lived, yet different vintages have far different characteristics and efficiencies. For example, equipment used today for switching includes mechanical cross-bar devices first installed in the 1920s and electronic components first introduced a few years ago. However, Solow [57] has demonstrated that even if the plant technology is one of fixed proportions, a Cobb-Douglas form may frequently give a good approximation to the long-run equilibrium conditions of the industry.

These input requirements functions relate inputs or components of cost to the level of output (with the implicit hypothesis that variations in factor prices are absent or are less important than variations in output for the determination of input levels). They may therefore be thought of simply as decompositions of conventional cost curves.

Such functions may be estimated in the form

$$I_{it} = C_i + \alpha_i Q_t + \beta_i M_t + \nu_i D_t + u_i \qquad (19)$$

where

I_{it} denotes the flow of input i in period t,
Q_t denotes revenue in period t,
M_t denotes the number of messages in period t,[22] and
D_t denotes an index of technical change at time t.

If there are economies of scale in the input i, one would expect the estimated constant term C_i in the above equation to be positive and statistically significant.

While input requirements functions focus on the question of economies of scale in particular inputs by studying functions relating inputs or costs, one by one, to levels of output, an alternative approach focuses on the issue of factor substitutability by studying functions relating input proportions to input price ratios. Such functions we may call productivity functions. In these, rather than explaining the demand for an input as a function of output and relative prices, we estimate the ratio of the input to output as a function of relative prices (both current and lagged). This procedure accords directly with the logic of the cost-minimizing model; moreover, the ratios of inputs to output (or their inverses) have traditionally been of interest. The ratio of output to labor is average labor productivity, while the ratio of capital to output is a statistic often used in analyzing economic growth. Yet a third ratio of interest is the capital/labor ratio, and this too we have attempted to explain.

For these aggregates, the functions estimated have been of the form

$$(Q/I)_t^i = \alpha(w/pr)_t^i + \beta(w/pr)_{t-1}^i + u_i, \qquad (20)$$

where Q represents an output measure, I an input measure, w an average wage rate, r an interest rate or rate of return on capital, and p a price of investment goods.

Estimation of these functions requires, strictly speaking, data on physical inputs, physical output, factor prices, and changes in technology. Moroever, the data are required for the operations of a single plant over time (or for a number of different-sized plants of the same vintage, at the same time). These ideal data are not available to us for this industry. First, the ideal measure of physical output for estimating the productive characteristics of telecommunications would require identifying as separate products all of the outputs which entail different inputs. A call from New York to Boston does not use the same inputs as a call from Los Angeles to San Francisco. Nor do two calls between two points necessarily use the same plant.[23] The ideal physical output measure might then involve specification of origin, destination, time, and duration: thus every call is a separate output!

---

[22] Both revenue and the number of messages are included in these equations because of the very different impact which increases in demand for local service, as compared to long-distance services, have on input requirements in the short run. Since local service is sold for a fixed fee, if the demand for local calls increases while the demand for the number of telephones remains constant, there are increased input requirements but no increase in revenue. When the output measure is local revenue, a short-run increase in telephone calls actually may lead to a decrease in measured labor productivity.

We tried two means of aggregating outputs; one was to use the number of messages, the other to use constant-dollar revenue. For estimating production relationships one might expect that the number of messages would be more appropriate, since most factor inputs into the production of telecommunications services would seem to be required for setting up a circuit rather than for maintaining the connection. Operator-handling or -switching activities certainly have this feature. Yet our tests using messages as the dependent variable yielded equations substantially inferior to those using revenue. On the other hand, since it is obvious that imperfect competition exists in the product market, our use of value rather than physical output data may lead to inconsistent and biased estimation (where the estimated parameters of the production function include measures of the elasticity of demand as well as the true parameters).[24]

In addition to the bias possible because of the use of data on value as measures of output, another possible bias in the estimation of an industry production function exists because of differences in technology at any point in time among firms. It can be shown[25] that an industry production function aggregating over firms is theoretically justified if, and only if, capital differs between firms by a capital-augmenting factor. An analysis to determine, along the lines of a procedure established by Solow,[26] the nature of the technological change implicit in the growth of Bell Canada suggested the presence of capital-augmenting technological change. Assuming that the other firms in the industry adopted changes in technology similar to those adopted by Bell Canada, but perhaps more slowly, it may be legitimate to represent firms as differing by a capital-agumenting factor, and hence to propose an industry production function.

Two further problems must be faced before we turn to the actual estimation: first, our inability to separate the different activities of the firm; and second, the effects of rate-base regulation on the estimation procedures.

The enterprise which generates output might be subdivided into three major activities, as illustrated in Figure 1: local collecting and distributing (consumer loops), switching, and trunking. A different role and different technology are involved in each of these.

Collecting involves the terminal equipment—telephones and teletypes or other peripheral equipment—necessary to relay messages

---

[23] A call between New York and Boston at the morning peak may be relayed via Los Angeles, because time zone differences mean that the routes from New York to Los Angeles and from Los Angeles to Boston are underutilized.

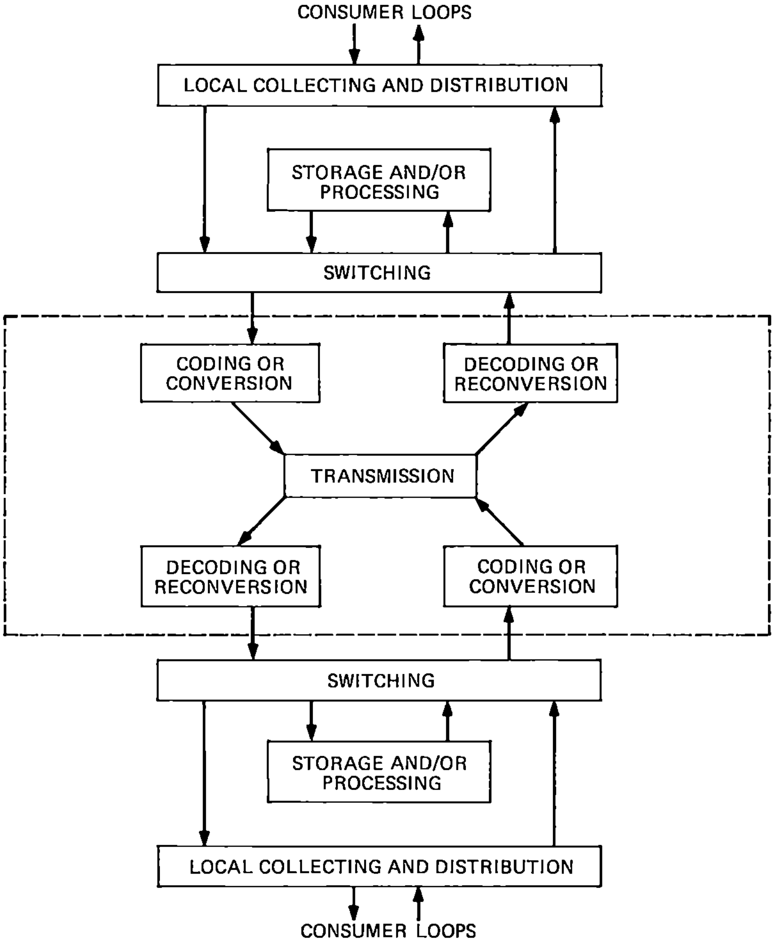
[24] See [38], [45], and [51].

[25] See [26].

[26] See [56].

FIGURE 1

ACTIVITIES INVOLVED IN PROVIDING TELECOMMUNICATIONS SERVICE, SCHEMATIC DIAGRAM



to, or collect messages from, individual callers, and the local wire loops connecting the individual to a local switching center. There, switching equipment routes local calls to the appropriate local exchange and toll calls to a nearby toll center, at which point the call is routed to an appropriate trunk line. Trunking then is the process by which switched messages are transmitted to a geographically distant toll center by means of wire line, coaxial cable, or microwave facilities.

Not all of the companies in the sector provide these three activities to the same degree. While the eight members of TCTS maintain large trunking systems, most of the other telephone companies provide only local loops and local switching stations. It would clearly be uneconomic for each company (including those whose subscribers are concentrated in one small geographic area) to provide the trunks whereby its subscribers could connect with any person in Canada. On the contrary, it is efficient for geographically limited companies to rent trunking services as they are needed from the large telephone systems.

Because of the joint use of common plant by most telecommunications services, estimates of the contribution of factor inputs to each activity separately are impossible to obtain from aggregate data.

Long-distance voice messages, for example, involve the facilities of all three activities. It is impossible to separate equipment by type of output (at least within present classifications); the same problem exists, moreover, with respect to labor inputs. For a single company, therefore, the type of function one can estimate with the data presently available must involve aggregate revenue (or messages) as the dependent variable, with total capital and aggregate labor inputs for the firm as the primary explanatory variables.

Despite the fact that traditionally, in industry, the plant is considered the basic unit in production, in telecommunications our specification must focus on the productive characteristics of the firm. The reason is simply that, although it is in the plant that technological economies in combining and enlarging related activities are usually assumed to be realized, the distinction between firm and plant is blurred in the telecommunications industry. The geographic isolation which helps to define "the plant" for manufacturing industries does not aid us in this sector, for each activity is, by its very nature, geographically dispersed. As a result, statistical estimates of productive characteristics can only be attempted for firms. One immediate consequence of this limitation is that we are unable to analyze possible economies of scale within the separate activities, even though such economies might arise in any of the three major activities—collecting, switching, and trunking—or at the level of the firm.

A similar difficulty arises in attempting to compare results for separate firms or subsectors of the industry. Enterprises in the industry offer different mixes of services and face different cost curves within their separate activities; this heterogeneity of services means that neither the production functions nor the input requirements functions for different subsectors of the industry need be comparable. The upshot of this is that characteristics such as the contribution to production of different inputs or the importance of economies of scale may depend heavily on the specific activities of the particular company analyzed.

Finally, the effect of rate-base regulation on the input and output choices of the public utility is a fascinating issue, discussed at length in recent literature[27] but still worth much more investigation than is possible here. We take the position that in the estimation of production or input requirements functions, which involve (in principle) only technical, not financial, optimization, possible distortions introduced by regulation need not bias the estimation procedure. Whether or not the level of durable physical capital put in place by the firm is unduly large in the light of factor prices prevailing on the market, as long as the firm employs the minimum labor inputs necessary with the fixed capital to produce the observed output, the production function could be estimated. Difficulties arise, however, with any attempt to estimate productivity functions, which do presume explicit cost minimization by the firm. As will be seen, the productivity functions estimated from the available data are somewhat inconsistent with the hypothesis of cost minimization under the constraint of the Cobb-Douglas production function estimated for the same period.

---

[27] See [60] or [5], for example.

Thus there are several conceptual difficulties in the analysis of production relationships within the sector: (1) joint use of common plant prevents identification of separate relations within the separate activities of the sector; (2) lack of physical measures forces estimation of relationships between constant dollar magnitudes; and (3) engineering data are too detailed for the aggregate relationships we seek.

In addition, a number of practical limitations on available data have to be noted. In general, we have no data on inputs of capital services. What we have are data on a capital stock (and frequently only a book value, sometimes depreciated, of that capital stock). Thus, at best, we require a capacity-utilization index which would allow the conversion of the capital-in-place series to a capital services series. No such series suggests itself, however. Conventional approaches to constructing a capacity-utilization index obviously fail when applied to a network with a dramatic peak-load problem and a geographical extent which ensures that much of the network must be idle much of the time. Seeing no alternative, therefore, we employ data on deflated value of plant, book value of plant net of accumulated depreciation, or book value of plant gross of accumulated depreciation as substitutes for data on inputs of capital services.

Similarly, except in the case of data for Bell Canada, we have used numbers of employees as a measure of labor input, even though these values do not reflect any change in the number of hours worked per man per year or any changes in the skill mix within the labor force. Where possible—that is, in the relations estimated for Bell Canada data—we have employed a wage-weighted index of labor hours as a measure of labor services.

Thus, to summarize, our investigation of the characteristics of production in the telecommunications sector is based initially upon the use of an overall production function. In principle, this single function reflects all relevant aggregate features of the prevailing physical constraints upon the transformation of inputs into outputs. Subsequently we studied input requirements functions and productivity functions separately, the former relating inputs or components of costs to levels of output, the latter relating factor proportions to factor prices. In the following subsections, we describe the results obtained from estimates of such functions for Bell Canada and the telephone industry in general.

☐ **Empirical results for Bell Canada.** As indicated above, data availability varied among firms and subsectors of this industry, with data relating to a single firm—Bell Canada—being most complete and appropriate.

Bell Canada is the dominant firm in the telecommunications industry in Canada. In 1967, it earned 60.6 percent of all revenue earned by telephone companies and 55 percent of the revenue earned by telephone and telegraph companies together.[28] The next largest

---

[28] This measure considerably understates Bell Canada's actual size, for the company has substantial interest in other telephone companies whose revenue is not consolidated with that of Bell: New Brunswick Telephone, Avalon Telephone Company, Northern Telephone, and Maritime Telephone and Telegraph (and a number of smaller telephone companies). Bell also owns 100 percent of Northern Electric, Canada's largest telecommunications equipment manufacturer.

TABLE 1

PRODUCTION CHARACTERISTICS, BELL CANADA, 1952–1967

	REVENUE		CAPITAL		LABOR			RATIOS				
								CAP./LABOR			REV./LABOR	
YEAR	REVENUE (000)	PERCENT CHANGE IN REV., YR. TO YR.	NET STOCK (000)	PERCENT CHANGE IN K, YR TO YR.	NO.MEN (000)	NO MAN-HRS. (MILLIONS)	PERCENT CHANGE, MAN-HRS	K/NO.	K/HRS	CAP./REV.	REV./NO.	REV./HRS.
1952	$199.6		$564.9		28.7	44.9		19,685	12.58	2.83	6955	4.45
1953	216.0	8.2	620.9	19.9	30.0	45.8	2.0	20,700	13.56	2.87	7200	4.72
1954	234.7	8.7	689.4	11.1	31.1	47.4	3.5	22,165	14.54	2.98	7550	4.95
1955	260.6	11.1	782.8	13.5	34.6	50.9	7.4	22,625	15.38	3.00	7530	5.12
1956	289.7	11.4	889.6	13.6	37.8	54.4	6.9	23,535	16.35	3.07	7665	5.33
1957	320.2	10.5	1017.7	14.4	39.5	57.1	5.0	25,765	17.82	3.18	8105	5.61
1958	343.7	7.3	1167.5	14.7	37.2	57.8	1.2	31,390	20.20	3.40	9235	5.95
1959	373.0	8.5	1314.2	12.6	35.0	56.9	−1.6	37,550	23.10	3.52	10,655	6.56
1960	398.8	6.9	1463.5	11.4	34.1	55.2	−3.0	42,920	26.51	3.67	11,695	7.22
1961	426.7	7.0	1601.3	9.4	32.9	53.6	−4.7	48,670	29.88	3.75	12,910	7.96
1962	470.5	9.3	1734.0	8.2	33.6	52.8	−1.5	51,605	32.84	3.69	14,005	8.91
1963	503.1	6.9	1876.3	8.2	33.8	53.7	1.7	55,510	34.94	3.73	14,865	9.37
1964	542.5	7.8	2014.0	7.8	34.5	55.1	2.6	58,375	36.55	3.71	15,725	9.85
1965	593.1	9.3	2143.9	6.5	35.5	56.3	1.5	60,390	38.08	3.61	16,710	10.53
1966	648.3	9.3	2285.6	6.6	37.5	57.3	1.8	60,950	38.89	3.53	17,290	11.31
1967	705.0	8.8	2438.7	6.7	36.8	58.4	2.3	66,270	41.76	3.46	19,160	12.08

REVENUE—TOTAL FIRM REVENUE MEASURED IN CONSTANT 1967 DOLLARS.

LABOR — MAN—HOURS ADJUSTED FOR THE CHANGING MIX OF EMPLOYMENT (SEE NOTE 33 FOR AN EXPLANATION OF THE DERIVATION OF THIS SERIES)

CAPITAL — NET STOCK IN CONSTANT 1967 DOLLARS.

SOURCES NUMBER OF EMPLOYEES—DBS TELEPHONE STATISTICS [16], OTHER DATA — OLLEY [53]

telephone company, British Columbia Telephone, is only one-sixth the size of Bell Canada.

1. Production function

Table 1 shows that apparently there has been considerable change in the productive characteristics of Bell Canada. The capital/labor ratio measured in dollars of net capital stock per man-hour increased from 12.58 in 1952 to 41.75 in 1967 (a growth of 232 percent). The capital/output ratio increased from 2.83 in 1952 to 3.75 in 1961, then decreased to 3.46 in 1967. Revenue per man-hour increased 171 percent in this same period. The number of man-hours increased 30 percent between 1952 and 1967, although this figure masks a decrease in man-hours for the years 1959–1962. The capital stock, however, grew some 332 percent over the same period.

An analysis of the annual rates of change of the revenue, labor, and capital series demonstrates marked differences in the within-decade growth of each series between the 1950s and the 1960s. The annual percentage increases in the measured capital stock averaged 17.7 percent in the 1952–1960 period and 9.5 percent in the period

TABLE 2

PERCENTAGE OF STATION–TO–STATION TOLL CALLS DIALED BY THE CUSTOMER

YEAR	BELL	MARITIME	NEW BRUNSWICK	QUEBEC (EXCL. BELL)	MANITOBA	SASKATCH-EWAN	ALBERTA	BRITISH COLUMBIA	OKANAGAN
1952	0	0	0	0	0	0	0	0	0
1953	0	0	0	0	0	0	0	0	0
1954	0	0	0	0	0	0	0	0	0
1955	0	0	0	0	0	0	0	0	0
1956	0	0	0	0	0	0	0	0	0
1957	1.4	0	0	0	0	0	0	0	0
1958	6.0	0	0	0	0	0	0	0	0
1959	10.3	0	0	0	0	0	0	0	0
1960	17.3	0	0	0	0	0	0	0	21.0
1961	23.1	0	0	0	0	0	0.9	11.3	53.0
1962	27.0	0	0	0	0	0	2.7	23.4	55.0
1963	31.9	0	0	0	0	0	5.0	35.5	55.0
1964	38.3	10.5	1.2	0	0	7.3	14.5	38.7	56.0
1965	44.1	11.7	12.3	0	11.7	13.2	21.5	39.0	54.0
1966	47.9	12.0	20.3	11.2	16.2	19.6	28.2	39.8	52.0
1967	51.6	22.3	36.5	15.4	20.3	28.1	42.6	39.4	54.0

SOURCE: DATA PROVIDED BY TRANS–CANADA TELEPHONE ASSOCIATION.

1961–1967. In the years 1958–1962 while the capital stock increased nearly 40 percent, the number of man-hours worked decreased by 8.7 percent.

A possible explanation of these trends is not difficult to find.[29]

> Direct Distance Dialing was first introduced in Canada in 1956. . . . In 1958, the first full-scale Canadian installation was completed in Toronto. . . . When dial equipment first took over the job of handling local telephone calls, there were predictions that thousands of operators and other telephone people would be out of work. To the contrary, after a temporary halt in the employment curve . . . improved transmission methods are generating greater demands for service. . . .[30]

Thus the marked differences which appear in the data after 1958 may be due to the heavy capital investment undertaken by Bell Canada prior to 1958 in order to implement a more efficient long-

---

[29] Refer to [61].

[30] Further evidence of the significance of the 1958 break can be offered. A regression of the form $R = AL^{\alpha}K^{\beta}e^{\gamma t}$ for the period 1952–1967 yielded poor results, with high serial correlation in the residuals. Analysis of the residuals of this equation showed that between 1952 and 1957 the residuals are positive, from 1959 to 1964 they are negative, and for 1965 through 1967 they are once again positive. Output predicted from the combinations of labor and capital is thus being overestimated in the periods 1952–1957 and 1964–1967 and underestimated in the middle period. A division of the period into two subperiods, 1952–1958 (direct distance dialing came to Ontario late in 1958) and 1959–1967, yielded far better results, with the serial correlation eliminated. Covariance analysis indicated that the subperiod regressions did in fact come from different populations.

TABLE 3

CONSTANT–DOLLAR (1967) RAW MATERIALS AND INDIRECT TAXES AS PERCENTAGES OF CONSTANT–DOLLAR (1967) REVENUE

YEAR	RAW MATERIALS, SERVICES, RENT, AND SUPPLIES	INDIRECT TAXES	SUM (COLUMNS 2 AND 3)
1952	20.1	4.6	24.7
1953	19.9	4.6	24.5
1954	20.3	4.6	24.9
1955	21.0	4.6	25.6
1956	22.1	4.5	26.6
1957	20.0	4.6	24.6
1958	20.4	4.6	25.0
1959	19.6	4.7	24.3
1960	19.1	4.9	24.0
1961	18.6	5.2	23.8
1962	16.9	5.0	21.9
1963	17.8	4.9	22.7
1964	16.5	4.8	21.3
1965	16.5	4.7	21.3
1966	15.7	4.9	20.6
1967	14.0	5.0	19.0

SOURCE: OLLEY [53]

distance calling service.[31] This investment was not fully completed until 1958, when customers in Toronto could call other areas by direct distance dialing (DDD). The successful implementation of DDD in that year led to a reduction in employment in the three following years and a generally lower rate of growth of labor services in the 1960s.[32]

Table 2 gives the percentage of station-to-station toll calls dialed by the customer for ten companies during the period 1952–1967. The rise in this percentage from zero reflects the technological change discussed above—namely, the replacement of operator-switching systems by machine-switching systems. This percentage of station-to-station toll calls dialed by the customer has been used as a proxy for technological change.

Although this index implies the absence of significant technological progress over the period 1952–1957, it is the best measure of the pace of technological progress we found. Attempts to introduce an additional trend to account for improving technology over this early period failed, presumably because any improvements were masked by heavy investments not utilized within the period.

---

[31] We are not suggesting that this was the only technological advance in this period. For example, linemen became more efficient through the use of crane trucks (telsta trucks) which eliminated climbing of poles.

[32] Before direct distance dialing, an operator answered the call "by plugging one end of a cord circuit into a jack (or hole) on the switchboard; and completed the call by plugging the other end of the cord into an outgoing jack" [61], pp. 43–4.

To estimate the production function in value-added terms, we deduct from the dependent variable (1) the amounts of raw materials required by the production process, and (2) the indirect taxes paid by Bell Canada. Table 3 gives the percentage of constant-dollar revenue accounted for by the intermediate inputs of raw materials, services, rent, and supplies. The indirect taxes paid by Bell Canada are also shown as a percentage of constant-dollar revenue. As can be seen from the table, raw material requirements fell in this period from 20.1 percent of revenue in 1952 to 14 percent of revenue in 1967, while indirect taxes were a relatively constant percentage in this period.

The Cobb-Douglas function estimated for Bell Canada is therefore of the form

$$V_t = AL_t^{\alpha}K_t^{\beta}e^{\gamma D_t} \tag{21a}$$

or

$$\log V_t = \log A + \alpha \log L_t + \beta \log K_t + \gamma D_t, \tag{21b}$$

where

V_t = constant-dollar value added, year t (1967 dollars, millions),

L_t = man-hours in year t, adjusted for the changing mix of employment (millions),

K_t = net capital stock year t, adjusted for real depreciation, (1967 dollars, millions), and

D_t = percentage of station-to-station toll calls dialed by the customer.[33]

The equation is

$$\begin{array}{cccccc} \log V_t = & \log A + & \alpha \log L_t + & \beta \log K_t + & \gamma D_t & \\ & -0.25 & 0.705 & 0.405 & 0.010 & \quad (21c) \\ & (-0.81) & (4.9) & (6.8) & (9.5) & \end{array}$$

with

$$R^2 = 0.998, \quad DW = 1.78, \quad \text{and} \quad S_e = 0.022.$$

Statistically this equation appears to be thoroughly satisfactory. The coefficient of the technological-change term has the appropriate sign and indicates that an increase in the percentage of toll calls handled by the customer permits the handling of a larger number of

---

[33] The source for data on V,L,K was Olley [53]. A brief description of the procedure used by him to arrive at real-dollar estimates is as follows:

R = current dollar series, adjusted by Bell index of price changes for subcategories of revenue.

L = actual hours worked, adjusted for skill differentials by assuming that the wage for each type of labor measured its skill in relation to the average wage for all labor in the base year 1967. For example, if the average wage for all employees was \$3.00/hour, and telephone operators received \$2.00/hour, then it is assumed that operators are two-thirds as skilled as the average laborer and the total hours worked by operators is multiplied by 0.667.

K = a net constant dollar capital stock series constructed using telephone plant price indexes (1967 base) and age distributions of the net stocks of physical capital in place in each year.

The data used in estimation are marginally different from the data presented in the preceding tables. The regressions utilize data presented by Olley in [54] rather than [53]. The variable V was obtained by subtracting from constant-dollar revenue, R, estimates of outlays for materials and direct taxes as described above.

TABLE 4a

INPUT REQUIREMENTS FUNCTIONS USING REVENUES AND MESSAGES

I_{it}	$= C$	$+ aQ_t$	$+ bM_t$	$+ cD_t$	$\bar{R}^2$	DW
(i) TELEPHONE OPERATORS	115.3 (6 9)	0.238 (3 .1)	–0.016 (–2 5)	–1.43 (–3.3)	0 9176	0.82
(ii) PLANT CRAFTSMEN	9 54 (0 63)	0.156 (2 3)	0 019 (3.2)	–2.11 (–5 4)	0.9339	1.31
(iii) RAW MATERIALS, (LOG)	2.52 (16 0)	0 0006 (0.91)	0 0002 (5 8)	–0 022* (–5 3)	0 9729	1.54
(iv) NUMBER OF EMPLOYEES (LOG)	2.84 (29 0)	0.002 (4.7)	0.00004 (1 0)	–0 019 (–7 3)	0 8597	1 15

*TIME TREND RATHER THAN D_t.

TELEPHONE OPERATORS = MILLIONS OF HOURS (ADJUSTED FOR CHANGES IN THE SKILL LEVEL)

PLANT CRAFTSMEN = MILLIONS OF HOURS (ADJUSTED)

RAW MATERIALS = MILLIONS OF CONSTANT (1967) DOLLARS

EMPLOYEES = NUMBER OF FULL–TIME EMPLOYEES, THOUSANDS

INDIRECT TAXES = MILLIONS OF CONSTANT (1967) DOLLARS

TELEPHONES = THOUSANDS OF TELEPHONES

Q_t = REVENUE, MILLIONS OF CONSTANT (1967) DOLLARS

M_t = MESSAGES, BILLIONS

D_t = PERCENTAGE OF TOLL STATION–TO–STATION CALLS DIALED BY THE CUSTOMER

SOURCES WAGES, PLANT CRAFTSMEN HOURS, OPERATOR HOURS,
NUMBER OF TELEPHONES,
NUMBER OF MESSAGES – BELL CANADA DATA.
OTHER VARIABLES–OLLEY [53]

calls (and hence increases revenue) without corresponding increases in factor inputs. The coefficient on the variable D_t represents an estimated rate of technical progress of about 4.75 percent per year, after completion of the initial gestation lag on the heavy initial investment. Its inclusion in the equation substantially improves the results: the percentage of explained variation rises, and the degree of serial correlation falls. The sum $\alpha + \beta$ of the coefficients of capital and labor (the measure of the economies-of-scale parameter) is 1.110, which is significantly greater than unity at the 97.6 percent level and implies that a doubling of all inputs would raise output by 110 percent rather than 100 percent as would be the case with a constant returns-to-scale function. Production function estimates for Bell Canada employing either alternative formulations or alternative data series are available from the authors on request.

In order to obtain a somewhat different perspective on the question of returns to scale, we also estimated input requirements functions of the form indicated above. Let us now turn to a discussion of these.

2. *Input requirements (cost) functions*

Estimating a relationship between output or scale of operation (as measured by revenues and numbers of messages) and various inputs

TABLE 4b

INPUT REQUIREMENTS FUNCTIONS USING REVENUES ONLY

I_{it} =	C +	aQ_t +	bD_t	$\bar{R}^2$	DW
(i) WAGES	−12 5 (−1 5)	0 524 (17.0)	−2 25 (−8 7)	0 997	2.21
(ii) NUMBER OF EMPLOYEES (LOG)	2.90 (38 0)	0.002 (8 1)	−0.018 (−7 4)	0.8479	1.17
(iii) NUMBER OF MAN-HOURS (ADJUSTED)	24.15 (6 8)	0.111 (8 0)	−0.738 (−6.5)	0.8993	1.14
(iv) TELEPHONE OPERATORS (ADJUSTED HOURS)	89 5 (5 8)	0.103 (1 7)	−1.75 (−3 6)	0 8763	0 53
(v) PLANT CRAFTSMEN (LOG) (ADJUSTED HOURS)	0.329 (1.2)	0 799* (17 0)	−0 010 (−9.7)	0 9808	1 98
(vi) RAW MATERIALS	45.78 (18 0)	−0.073 (−3.1)	6.65** (8.5)	0.9899	1 93

*LOG OF REVENUE

**TIME TREND RATHER THAN D_t

WAGES = TOTAL WAGE BILL DEFLATED BY BELL REVENUE / PRICE INDEX (MILLIONS OF 1967 DOLLARS)

EMPLOYEES = NUMBER OF FULL TIME EMPLOYEES, THOUSANDS (OTHER VARIABLES AS DEFINED IN TABLE 4a)

SOURCES NUMBER OF EMPLOYEES – BELL CANADA DATA (OTHER VARIABLES AS GIVEN IN TABLE 4a).

or components of cost, we obtain the results presented in Table 4a. As indicated above, we look to the sign and the significance of the intercept term for an indication of the presence or absence of increasing returns to scale. However, the possible multicollinearity introduced by the simultaneous inclusion of two output variables (which generally must change together) may bias the coefficients; therefore we present in Table 4b estimates of the relationship between various inputs and a single measure of output, namely revenue in constant dollars.

In Table 4a, three of the four equations display significantly positive constants. Equation (ii) suggests that economies of scale do not occur in the utilization of the services for plant craftsmen who install and repair equipment. Of the six equations in Table 4b, four exhibit increasing returns: the average number of employees, the average number of adjusted man-hours, the average adjusted hours of telephone operators, and average raw material inputs all decrease with increased output. Hours of plant craftsmen per dollar of revenue and wages per dollar of revenue appear to be unrelated to output scale.

Equation (i), Table 4a, suggests that an additional dollar of revenue is linked with 0.24 of an operator-hour, while an increase in the number of messages reduces labor requirements, as does an increase in the percentage of toll calls dialed by the customer. The negative relation between messages and operators is difficult to explain, except as a consequence of the interaction between the two measures of output included in the equation. An additional dollar of revenue is associated with 0.16 hours of plant craftsmen's time, while an additional message is accompanied by 0.02 hours of such time. The negative correlation between plant craftsmen and the percentage of calls dialed by customers may reflect the decreased maintenance which more modern equipment requires as compared with older plug-in type switchboards. Raw materials [equation (vi), Table 4b] are negatively related to output measures. This result suggests that raw materials represent an overhead input for this sector, for, as we have observed before, much of material inputs apparently flows to construction and maintenance.

3. *Productivity functions*

Cost curves of the kind derived above take output to be the primary determinant of input requirements, omitting mention of factor costs. In theory, such requirements functions describe the results obtained by comparing the costs associated with various hypothetical levels of output under the assumption that input prices are unchanged and independent of output levels.

The introduction to this section suggested that the regulated firm also adjusts its capital/labor ratio in response to changes in factor prices. We therefore attempted to derive, for the period 1952–1967, a set of relationships relating both the capital/labor and output/labor (labor productivity) ratios to changes in factor prices.

Table 5 presents the results of our attempts at estimating productivity functions. The first two equations show the response of the capital/labor ratio to the wage/rental ratio in the same time period, the wage/rental ratio lagged one period, and a technology term. As can be seen in the first equation, the coefficient of the present period factor price ratio (*a*) is insignificant. This is not surprising since one

TABLE 5

ESTIMATED PRODUCTIVITY FUNCTIONS—BELL CANADA

	$\bar{R}^2$	DW
(i) $\mathrm{LOG}\ (K/L)_t = C + a\,\mathrm{LOG}\left(\frac{w}{r}\right)_t + b\,\mathrm{LOG}\left(\frac{w}{r}\right)_{t-1} + c\,D_t$ 1 75 0.13 0.23 0.015 (4 8) (1.2) (2.1) (6.5)	0.9906	0.9034
(ii) $\mathrm{LOG}\ (K/L)_t = C + b\,\mathrm{LOG}\left(\frac{w}{r}\right)_{t-1} + c\,D_t$ 1.98 0.28 0 015 (6 2) (2.7) (6 6)	0 9883	1.20
(iii) $\mathrm{LOG}\ (V/L)_t = C + b\,\mathrm{LOG}\left(\frac{w}{r}\right)_{t-1} + c\,D_t$ 0.66 0.23 0 015 (5.0) (5.2) (28 2)	0 9938	1.38

$(K/L)_t$ = CAPITAL / LABOR RATIO, TIME t

$(V/L)_t$ = VALUE ADDED PER HOUR WORKED, TIME t

w_t = AVERAGE WAGE RATE, TIME t

r_t = PRICE OF CAPITAL SERVICES, TIME t (CALCULATED AS THE RATE OF RETURN ON TOTAL CAPITAL MULTIPLIED BY AN INDEX OF THE COST OF CAPITAL GOODS)

D_t = PERCENTAGE OF CALLS DIALED BY THE CUSTOMER, TIME t

SOURCES D_t, RATE OF RETURN ON TOTAL CAPITAL—BELL CANADA,
ALL OTHER SERIES—OLLEY [53]

would hardly expect that an industry with a large amount of fixed capital would immediately change its use of factors in response to a change in the relative prices of these inputs. The coefficients in all three equations have the right signs and are all highly significant (except for the current relative factor price term in the first equation). The Durbin-Watson statistics suggest the presence of some serial correlation.[34]

For the first two equations in Table 5, the sum of the coefficients *a* and *b* is an estimate of the elasticity of factor susbtitution. With an underlying Cobb-Douglas function, one expects this elasticity of factor substitution to approximate unity. Our estimates of this elasticity are significantly less than unity and thus are inconsistent with the production function estimated earlier.

The third equation in Table 5 indicates that an increase in the wage/rental ratio increases labor productivity; capital is substituted for labor, and average labor productivity increases as a result. Again, this elasticity is significantly less than one, consistent with the findings for the elasticity of response in the capital/labor ratio with respect to relative factor price changes.

We may summarize our results from the analysis of production data for Bell Canada by observing that:

(1) An index of technological change based upon the percentage of station-to-station toll calls dialed by the customer plays a crucial role in explaining apparently disparate movements in inputs and outputs over the period 1952–1968.

(2) Use of this index, together with data on deflated values of revenue and plant and wage-weighted labor input measures, enables estimation of a statistically satisfactory aggregate production function of Cobb-Douglas form.

(3) This production function suggests the presence of modest increasing returns to scale over Bell's operations as a whole, although these cannot be isolated within any specific sub-activity.

(4) Estimation of more detailed input requirements functions suggests that these economies of scale are not to be found in the total wage bill, even though average hours of labor input (except plant craftsmen) fall with increasing output but are accounted for in part by falling raw materials requirements per unit of output.

(5) Estimation of productivity functions suggests that the regulated firm does not respond to changes in factor prices with as great an elasticity as might an unregulated firm.

☐ **Empirical results for TCTS.**

1. Production function

Sufficient data were available for ten companies, either members or affiliates of TCTS, to estimate cross section/time series production functions for the years 1952–1967.[35] The results for two such equations—one of which includes Bell Canada, the other of which does not—are similar to each other but different from the results obtained in equation (21c) for Bell Canada alone. In particular, where the Bell Canada equation shows increasing returns to scale, both of the TCTS production functions exhibit constant returns to scale. Moreover, the coefficient on labor (α) is substantially greater in equation (21c) than in the TCTS equations.

---

[34] Preliminary estimates of the first two equations suggested a substantial degree of positive serial correlation in the residuals. The equations presented here utilized a method of eliminating such serial correlation by transforming the data according to the estimate of the degree of first-order serial correlation [32].

This marked and somewhat implausible difference in the results obtained from estimating production functions for Bell Canada and for member firms of TCTS can perhaps be explained solely by the substantial differences in the reliability of the data as between the TCTS regressions and the Bell Canada equation.[36] First, the latter used as its estimate for labor services a measured total man-hours index adjusted for the changing skill mix of the labor inputs; the former used the unadjusted number of full-time workers. Because of variations in the workweek over this period and the changing mix of skill present in the labor force, the raw employee data are inferior to the adjusted man-hours series as a measure of labor services. Second, in estimating the production function for Bell Canada alone, a better measure of net capital in constant dollar value was used than in the TCTS equations, where the gross book value of capital adjusted by a price series for fixed plant obtained from Bell was employed. Production functions estimated for Bell Canada using number of employees as the measure of labor services and deflated book value of capital as the measure of capital services in fact showed constant returns to scale.[37]

Thus, while the sum of $\alpha + \beta$ in the estimation of a production function for the ten companies is not greater than one, it is possible that this sum is an underestimate of the true economies-of-scale measure. Evidence to this effect is found in the fact that the economies-of-scale estimate for Bell Canada alone is lowered below our earlier estimate by substitution of the less adequate choice of independent variables to which we were restricted by TCTS data. It is therefore possible that slight increasing returns to scale typify the growth of the large telephone companies—as appears to be the case for Bell Canada—in the period 1952–1967. For further evidence, we again attempt the estimation of input requirements functions.

2. *Input requirements functions*

The DBS publication *Telephone Statistics*[38] divides expenses by province into six categories: depreciation, commercial, indirect taxes, traffic, repairs, and other expenses. Each of these six categories can be considered a form of input, and requirements functions can be estimated along the lines of the procedures described in the previous section. The simplest explanation for each of these categories was attempted by considering the input involved to be a function of either revenue or the capital stock (or both).[39]

---

[35] The companies are Bell Canada, Northern Telephone System, Quebec Telephone, Maritime Telephone and Telegraph, New Brunswick Telephone, Manitoba Telephone, Saskatchewan Telephone, Alberta Telephone, British Columbia Telephone, and Okanagan Telephone.

[36] The possibility that these differences in results might be explained in part by the statistical problem of heteroscedasticity has been checked. The residuals of the TCTS equations do not suggest evidence of this statistical difficulty.

[37] Information on these equations, not reported in detail in this paper, is available from the authors on request.

[38] See [16].

All categories, with the exception of depreciation, are highly correlated with revenue. Associated with an additional dollar of revenue are $0.081 of commercial cost, $0.046 of traffic costs, $0.080 of indirect taxes, $0.162 of repairs and $0.106 of other expenses. Commercial expense is negatively related to the capital stock, and indirect taxes is the only category with a negative trend. From the fact that of the intercept terms only that for depreciation is significant, one would conclude that these expense items are simply proportional to output, with no spreading of any overhead component except that on the plant itself.

3. Productivity functions

Results on productivity were obtained from estimates of functions relating the capital/labor ratio and the output/labor ratio to ratios of factor prices, the inverse of a measure of size, and a number of time dummies. These formulations were used for several reasons. Lagged relative factor prices series are not included because of the resulting loss in an already small number of observations. The measure of size (revenue in the equation explaining the capital/labor ratio, capital in the equation explaining the output/labor ratio) was intended to represent the differences in technology between provinces, which were a function of size. It was felt that the larger provinces had available a number of techniques (automatic switching and accounting, etc.) which significantly affected both the capital/labor ratio and labor productivity.

The signs on all the coefficients were as expected (except, of course, for the time dummies, where no *a priori* judgment was involved). Increases in the wage/rental ratio increase both the capital/labor ratio and labor productivity. In addition, the larger the province, the greater both the capital/labor and output/labor ratios. The coefficient which measures the elasticity of factor substitution was not significantly different from unity, the value predicted by an underlying Cobb-Douglas production function. Although this result is substantially different from that obtained with Bell Canada data (Table 5), the difference may again be attributable to the significant differences in the data and time periods underlying the two results.

Thus, in summary, our analysis of the data for TCTS as a whole, while not inconsistent with the earlier analysis for Bell Canada alone (because of differences in the data series available for regression analysis) suggests no direct evidence for the presence of economies of scale, either in the production function or in the more detailed input requirements functions. Nevertheless, with the introduction of regular shifts over time (which displace our earlier index of technological change), it was possible to estimate a cross section function explaining differences in company revenues on the basis of capital and labor inputs alone. Substantial regularities in production relationships across the different companies are thus suggested.

---

[39] A table showing estimated requirements functions is available from the authors.

☐ **Discussion of results.** What does the foregoing tell us about the characteristics of production in the telecommunications sector? A few strong features are evident.

First, the sector produces services which cannot be stockpiled or held in inventory, so that peak-load problems are severe, and lack of adequate capacity means unserviced demand and foregone revenue. None of our analyses and no available data enable us to assess the costs associated with the distribution of transmitted messages over the day or over the year, although it is obvious that price structures or the introduction of services bringing about a smoother distribution of messages must lower unit costs by spreading overhead plant and labor costs.

Second, since the network by which the sector has served its customers involves connections to each household or firm, the sector operates at least in part under conditions of "natural monopoly," meaning that a single firm can offer service at a lower cost at any relevant level of output than could any combination of separate firms.

It must be emphasized that the question of "natural monopoly" is not identical to the question of returns to scale. With a high fixed charge, one may encounter increasing unit costs (that is, decreasing returns to scale) as output increases but still have average costs below those attainable by any competition dividing the market. However, our results, as will be noted below, do suggest some conclusions with respect to the issue of returns to scale in utilization of factor inputs.

The trend in the use of factor inputs has been to rising capital intensity, with both capital/labor and capital/output ratios rising over the period. This trend is of course consistent with rising labor productivity and rising wage payments even when numbers of employees are not increasing rapidly. We can note four reasons which could account for this increase in capital intensity:

(1) Capital services becoming cheaper relative to labor,
(2) The shifting of demand toward services which can only be provided by more capital-intensive facilities,
(3) The lowest-cost production techniques being more capital-intensive at higher output levels, and
(4) The impact of rate-base regulation.

The first three reasons are really only one, of course: At the required scale of production, the most efficient way to meet the imposed demands involves introduction of increased amounts of equipment per employee.

Rate-base regulation can imply a rising capital intensity only if either the firm responds slowly over time to the presence of such regulation, adjusting its capital stock with lags to the regulated rate, or the divergence between the regulated rate and the cost of capital *narrows* over the period.[40] While rate-of-return data suggest that in fact the actual nominal return exceeded the constant allowed return in 11 of the 16 years in the sample period, we see no suggestion that the gap has in fact consistently narrowed. We therefore conclude that the presence of regulation cannot explain the tendency towards an increased capital intensity on the part of Bell Canada.

---

[40] This somewhat perverse result has led some people to question the validity of the model. In fact, since the model assumes a putty-putty technology and no

Our input requirements functions attempt to capture the impact of changing output levels on factor inputs in various categories. These, and the more elaborate production functions, do suggest the presence of some increasing returns to scale. In particular, the statistical relations which were estimated for Bell Canada for the period 1952–1967 suggest that the firm as a whole was operating under conditions of increasing returns to scale. The analyses of data for eight of the members of TCTS did not show increasing returns to be present. However, the data used in the latter analyses were not those appropriate to estimation of production of cost functions. In fact, when the functions for Bell Canada were rerun using variables to which we were limited in studying the other members of TCTS, the evidence of increasing returns to scale disappeared. This result suggests that greater confidence should be attached to the evidence from the equations estimated using Bell Canada data.

6. Investment

■ The physical plant of the Canadian telecommunications industry has expanded significantly over the period from 1950 to the present. Over this period, the average annual rate of growth in investment outlays (measured in constant dollars at annual rates) has been about 7 percent for the telephone industry and 9.2 percent for the telegraph and cable industry, compared with an average annual growth rate of 5.8 percent for business gross fixed capital formation in the economy as a whole. In 1950, the sector accounted for only 3.2 percent of total Canadian business investment in nonresidential construction, equipment, and machinery. This share rose to 6.6 percent in 1960 and to 7.1 percent in 1968. While there seems to be little reason to believe that such a rate of growth of capital investment in this industry could continue indefinitely in the future, expenditures on fixed capital stock by the sector may well outpace the rest of the Canadian economy for some time to come. Within the sector, the "telephone industry" accounts for about 95 percent of these outlays.

In this section investment functions for the Canadian telephone industry are described. Although many recent studies of investment behavior[41] conclude that the neoclassical model provides a better explanation of investment decisions than any of the competing models, we find that the level of investment expenditures by the Canadian telephone industry is best explained by a modified accelerator model. This result is consistent with one's expectations concerning the behavior of a regulated industry and is strongly supported by all our analysis of the available data.

In the following subsection, explanations are provided of data employed. Discussion of empirical results arising from various model

---

fixed capital, and does not explicitly include investment, its usefulness may be rather limited. Attempts to develop a more meaningful dynamic model have not made their way to the literature as yet. See [5].

[41] See [34], [35], and [37].

specifications will then be presented. A brief summary completes the discussion of investment decisions.

☐ **Data.** Three sets of data are needed to estimate conventional investment functions: (1) a series on investment expenditure by the industry, in constant dollars, (2) a series on stocks of capital as measured either in physical units of constant productivity or in constant dollars, and (3) a measure of the desired level of the capital stock.

Gross investment includes capital expenditures both on new construction and on equipment and machinery. To obtain a measure in physical terms, each series should be deflated by the appropriate investment-goods price index. Capital expenditure series adopted are estimated by DBS from its annual survey of private and public capital expenditures in Canada.

Price indexes of investment goods, prepared for their future report on the capital flows and stocks of the communications sector, were supplied directly by DBS. In some earlier estimates of the total investment function, an investment-goods deflator constructed from price indexes of goods that are either fully or chiefly manufactured, nonresidential construction materials, and wage rates in the construction industry was tried. There appear to be no significant differences in the final results employing the alternative deflators.

As in any production or investment study, the choice of a measure of capital stock presents special problems. Ideally, one would wish to measure capital stock by some physical capacity unit, but in empirical studies one could only hope to have a reasonable dollar measure. We have tried at least five different measures, none of which is altogether satisfactory. The measure that we have chosen was constructed, by a perpetual inventory method, directly from the 1926–1968 investment series in constant dollars, with a capital survival curve estimated by Bell Canada and an assumed rate of depreciation or obsolescence at 5 percent per annum.[42]

Output of the telephone industry is measured by the total operating revenue of the industry in constant dollars. There is no comprehensive revenue/price index for the whole industry; those employed here are from Bell Canada.

Market value of capital in the telephone industry is measured by the sum of market value of common shares and the long-term debt deflated by the GNP deflator. Market value of the common shares is obtained as the product of the book value of stock and the price index of common stock of the telephone industry. Both the book value of the common stock and the long-term debt are reported in *Telephone Statistics*. The price index of common stock is available in *Prices and Price Indexes*.[43]

---

[42] Setting the initial value of the capital stock in 1926 at zero, and employing the series from 1950 on, amounts to ignoring the value, *relative to current technological standards*, of all equipment more than 25 years old. The series is comparable to a Solow-type "vintage" capital stock measured relative to the shifting standard provided by current capacity units. Perhaps a preferred interpretation is that of a net capital stock series obtained by accounting first for actual mortality through the survival curve and then for depreciation or declining value through the exponential weighting factor.

[43] See [16] and [17], respectively.

We assume that the expected rate of profit is approximated by historically observed rates of return on capital. All the necessary data are reported in *Telephone Statistics.*[44] We also require a long-term utility bond yield. The series we have used is an annual average based on monthly data obtained from McLeod, Young and Weir Investment, Inc. A similar series is also obtained for a Bell Canada long-term bond yield.

To obtain a measure of the desired level of capital stock suitable for a "neoclassical model" we must calculate a "user-cost-of-capital service" from the price of investment goods (q), the tax structure, the cost of capital (r), and the rate of replacement (δ), as well as the rate of capital gain or loss ($\Delta q/q$), by the following equation:[45]

$$C = \frac{q}{1-u}\left[(1 - u\omega)\delta + r - \frac{\Delta q}{q}\right], \tag{22a}$$

where u is the corporation income tax rate and ω is the proportion of depreciation deductible for tax purposes. If capital gains are considered "transitory," then this user cost of capital may be written as:

$$C' = \frac{q}{1-u}\,[(1 - u\omega)\delta + r]\,. \tag{22b}$$

The corporation tax rate is the ratio between taxes and operating profit before taxes; the proportion of depreciation deductible is measured by the ratio of capital consumption allowances for the telephone industry[46] to current depreciation. Current depreciation of the telephone industry as a whole is estimated from current depreciation of the large telephone systems.[47]

There are at least six alternative measures of the cost of capital which are commonly employed in empirical studies.[48] The most appropriate one, according to Modigliani and Miller, is the long-term bond yield.[49]

A similar set of data is also available for Bell Canada, mostly from its annual reports, and the rest from Olley.[50] However, since in the investment series there is no breakdown into new equipment and new construction, it is not possible to estimate an equation for each of these components, nor does the investment expenditure series go back far enough to allow us to construct directly a capital stock as we were able to do for the industry as a whole.

☐ **Empirical results.** Applying the data discussed above for the period 1952–1967 to the various models summarized earlier, we found, not surprisingly, that the best model specification for both the telephone industry as a whole and for Bell Canada is the accelerator model. For each model, we have chosen the distributed lag function which yields the minimum standard error around the regression. The best equations of the accelerator model for the telephone industry as a whole are presented here:[51]

---

[44] See [16].

[45] See [34], [37].

[46] Reported in [13].

[47] Reported in [16].

[48] See [48].

[49] *Ibid.* In addition to this one, we have also tried the following three measures: (1) current earnings yields as measured by the ratio of profits after taxes to the market value of stocks; (2) profit rate as measured by the ratio of gross profits to the current value of capital; and (3) the ratio of book value of net worth to the market value of the firm, none of which is thoroughly satisfactory as a cost-of-capital measure.

[50] In [54].

$$IVTD_t = \underset{(3.6)}{107.885} + \underset{(5.0)}{2.238}\, DREVD_{t-1} + \underset{(2.1)}{0.362}\, DIVTD_{t-1} + \underset{(2.5)}{0.034}\, KT_{t-1} \qquad (23)$$

$$(R^2 = 0.9705, \quad S = 17.1149, \quad DW = 3.0184)$$

$$IVED_t = \underset{(2.9)}{48.839} + \underset{(5.7)}{1.809}\, DREVD_{t-1} + \underset{(2.0)}{0.355}\, DIVED_{t-1} + \underset{(2.1)}{0.034}\, KE_{t-1} \qquad (24)$$

$$(R^2 = 0.9749, \quad S = 12.0399, \quad DW = 2.4210)$$

$$IVCD_t = \underset{(2.7)}{61.834} + \underset{(1.7)}{0.472}\, DREVD_{t-1} + \underset{(1.4)}{0.355}\, DIVCD_{t-1} + \underset{(1.3)}{0.030}\, KC_{t-1}, \qquad (25)$$

$$(R^2 = 0.8073, \quad S = 11.3452, \quad DW = 2.1622)$$

where

$IVTD_t$ = capital expenditures on both new equipment and construction by the Canadian telephone industry (in millions of 1967 dollars),

$DIVTD_t$ = net investment calculated from an average rate of replacement implied by the total constructed capital stock KT_t,

KT_t = constructed capital stock (telephone plant) for the telephone industry as a whole,

$IVED_t$ = capital expenditures on equipment by the Canadian telephone industry (in millions of 1967 dollars),

$DIVED_t$ = net investment in equipment and machinery calculated from an average rate of replacement implied by KE_t,

KE_t = constructed capital stock (equipment and machinery) for the telephone industry as a whole,

$IVCD_t$ = capital expenditures on new construction by the Canadian telephone industry (in millions of 1967 dollars),

$DIVCD_t$ = net investment in new construction calculated from an average rate of replacement implied by KC_t,

---

[51] The presence of negative serial correlation suggests the possibility that the surprisingly low depreciation coefficient may be biased downward as a result of our model specifications failing to reflect completely the capital stock adjustment process, a problem of misspecification that may be difficult to avoid in an annual model. We leave this possible statistical problem aside.

KC_t = constructed capital stock (construction) for the telephone industry as a whole, and
$DREVD_t$ = change in the operating revenue of the Canadian telephone industry (in millions of 1967 dollars).

In such models, the response of investment to expected changes in the demand for capital services is distributed over time—that is, the response of investment, gross or net, hinges upon the time elapsed after the changes in the demand for capital services have been anticipated. The time patterns of response of investment to a permanent change in demand for capital services implied in the above equations are given in Table 6. The first column shows the time elapsed after a change in the demand for capital services. Columns 2 to 7 record the responses of net and gross investment corresponding to each of the three distributed lag functions.

Column 2, for example, gives the proportion of response of total net investment (equipment and construction) to a permanent change in demand for capital services. A change in demand for capital results in no change in investment for expansion until one year later. At the end of that period, about 63.8 percent of the required change in demand for capital will be completed; one year later, another 23.1 percent of the required change will be finished, and so forth. The bulk of the required change in net investment is therefore completed by the second year after a change in demand for capital has been anticipated. The entire investment project initiated in response to a change in the demand for capital is fully completed in the ninth

TABLE 6

INVESTMENT RESPONSE* (DISTRIBUTED–LAG PATTERN)

LAG	CALCULATED INVESTMENT					
	TOTAL		EQUIPMENT		CONSTRUCTION	
	NET	GROSS	NET	GROSS	NET	GROSS
0	0	0	0	0	0	0
1	0.6380	0.6380	0.6449	0.6449	0.6448	0.6448
2	0.2309	0.2526	0.2289	0.2510	0.2295	0.2486
3	0.0835	0.1132	0.0813	0.1109	0.0806	0.1076
4	0.0302	0.0627	0.0289	0.0614	0.0287	0.0574
5	0.0108	0.0445	0.0103	0.0439	0.0096	0.0396
6	0.0040	0.0376	0.0036	0.0374	0.0041	0.0328
7	0.0014	0.0353	0.0014	0.0353	0.0014	0.0314
8	0.0006	0.0345	0.0004	0.0346	0.0000	0.0301
9	0.0003	0.0342	0.0000	0.0342	0.0000	0.0000
10	0.0000	0.0339	0.0000	0.0342	0.0000	0.0000
11	0.0000	0.0339	0.0000	0.0339	0.0000	0.0000
12	0.0000	0.0339	0.0000	0.0339	0.0000	0.0000
13	0.0000	0.0339	0.0000	0.0339	0.0000	0.0000
14	0.0000	0.0339	0.0000	0.0339	0.0000	0.0000
15	0.0000	0.0339	0.0000	0.0339	0.0000	0.0000

* COMPUTED FROM DIFFERENCE EQUATION SOLUTION OF EQUATIONS (23), (24), AND (25)

year after the change has been anticipated; after that, net investment will become zero if no other change occurs.

The proportion of response of gross investment is recorded in the third column. As may be observed, the peak response (63.8 percent) is reached a year after the anticipated change in demand for capital services has been diagnosed; all of this expenditure is for net investment. After that, gross investment is the sum of net investment and replacement investment. Unlike the response of net investment, however, gross investment will not become zero; instead, it will eventually be equal to replacement charges on the incremental stock of capital required to meet the increase in demand.

It may be noticed that the time pattern of response of investment in both equipment and construction is the same as that for total investment. Although data used here are not completely comparable to those of other studies, the results seem to corroborate other findings. For example, in using U. S. quarterly data, Jorgenson[52] found that for all regulated industries as a whole there is no response of investment until the fifth quarter after the change in demand for capital services. Moreover, about 63.3 percent of the required change is completed by the eighth quarter.

In passing, we may also note that the capital/output ratio implied by the distributed-lag function of these investment functions is very close to the historical ratio. The steady-state coefficient of output calculated from the total investment equation is 3.508, compared with an average capital/output ratio of 3.7 for 1952–1967 from the DBS measure of capital stock. Even compared with the average capital/output ratio of 3.9 for the same period calculated from our constructed capital stock, this steady-state coefficient is encouragingly close, suggesting that the distributed lag function implied in the total investment function may be a good approximation to the actual investment process of the telephone industry.

Another interesting point is the consistency between the aggregate equation (1) and the disaggregated equations (2) and (3). Since gross investment on both new equipment and construction is the sum of capital outlays on its two components, one would expect that the corresponding coefficients of these three regressions are compatible. This indeed is the case. For example, the sum of the coefficients of the change in demand for capital services in equations (24) and (25) is 2.281; compared with 2.238, the coefficient from the aggregate equation, the discrepancy is less than 2 percent. Furthermore, the aggregated steady-state coefficient of output from the disaggregated equations is 3.537, which is less than one percent above the coefficient implied in the aggregate equation. This implies that a forecast of total investment expenditures using these disaggregated equations will not be very different from that using aggregate equations, but the distribution of these expenditures between the construction and communications equipment sectors is worth knowing.

Thus far we have only discussed an accelerator model. No doubt one would like to compare the relative performance of this model with all others. For our present purposes a brief summary is sufficient.[53]

---

[52] In [35].

[53] The results of all other model specifications and a detailed comparison of their performance are available from the authors.

In most cases, the accelerator model explains investment behavior better than other models, by more than 30 percent in terms of the standard error of estimate; it works slightly better than the expected-profit model, which in turn is superior to most of the Jorgensonian neoclassical models. Compared with the modified neoclassical models, which allow for separate coefficients on the change in output and the change in the relative price of output to the rental value of capital service, however, the accelerator model performs less well in terms of the standard error of estimate. This is hardly surprising since these models are extended versions of the accelerator model, with change in output as the principal explanatory variable and other variables as well.

Nevertheless, in the modified neoclassical models the coefficients corresponding to changes in the relative price either are insignificant at a 5-percent level of significance or are implausible in light of the theoretical model. So are the coefficients associated with changes in the desired capital stock, in all except one of the non-accelerator models. From this point of view, the accelerator model is overwhelmingly superior.

In addition to these two measures, the accelerator model also yields slightly better (or at least equally good) results from the point of view of the maximum number of significant coefficients associated with changes in the demand for capital services. All indications thus confirm that the straightforward accelerator model is likely to be the best we can do in any aggregate analysis of investment in the telecommunications sector.

Through the period 1950–1968, capital expenditures by Bell Canada accounted for nearly two-thirds of the capital outlays of the telephone industry; the proportion ranges from a low of 53 percent in 1953 to a high of 68 percent in 1961. It is therefore desirable to estimate the investment function for Bell Canada and to compare the results with those for the whole industry.

As expected, the best model specification for Bell Canada is again the accelerator model. The equation is

$$\begin{aligned} BIVTD = \underset{(2.4)}{51.713} &+ \underset{(3.7)}{2.386}\, DBREVD_{t-1} \\ &+ \underset{(3.0)}{0.585}\, DBIVTD_{t-2} + \underset{(1.5)}{0.023}\, BOK_{t-1}, \qquad (26) \end{aligned}$$

$$(R^2 = 0.9554, \quad S = 10.6073, \quad DW = 2.6717)$$

where

$BIVTD_t$ = capital expenditure of Bell Canada deflated by the Bell telephone plant deflator,

$DBREVD_t$ = the changes in Bell Canada's operating revenues deflated by Bell revenue deflators, and

$DBIVTD_t$ = net investment calculated from an average rate of replacement implied in Olley's measure of net capital stock for Bell Canada (BOK_t).

The results are generally in agreement with the corresponding equation for the telephone industry as a whole.

The Bell coefficient associated with the change in output is larger than that for the industry by about 5 percent. This may simply reflect the fact that Bell is growing relatively faster than the rest of the firms in this sector; with a rate of population growth in Ontario and Quebec above the national average, such a result would not be surprising. The estimated rate of replacement seems rather low. This result is somewhat unexpected since one might anticipate that, being the leader in this sector, Bell Canada would introduce technological changes at a faster rate to keep its leadership. However, one must be warned that since the capital measure for Bell is not constructed precisely as was our trial capital stock series the coefficients on the stock of capital in equations (23) and (26) are not strictly comparable; (nor have these equations made adequate allowance for the effects of technical change, in any case).

In comparing this accelerator model with other models from Bell Canada data, we again found it to be preferred, with the general relative performance of the equations being much the same as in the results for the telephone industry.

To sum up our results with respect to investment behavior, then, although many recent studies of investment decisions conclude that a "financial" or neoclassical model is the most relevant model specification for many industries, we have found that investment expenditure by the telecommunications sector in Canada is best explained by a modified accelerator model. This result is consistent with one's expectations concerning the behavior of a regulated industry for the reasons outlined earlier, and is strongly supported by all our analyses of the available data.

For the telephone segment of the industry as a whole, this model indicates that an increase in revenues of $1 million per annum would be associated with $2.2 million of capital expenditure in the following year, with a subsequent stream of capital outlays continuing until ultimately a total increase of $3.5 million in the capital stock is attained. The distribution of these outlays over time (very little response in the current year, about 63 percent of investment completed in the following year, and virtually all completed within two years) is significant and is consistent with estimates obtained from quarterly data on regulated industries in the U. S. The fact that the long-run or steady-state coefficient of 3.5 is very close to the observed average capital/output ratio for the industry suggests that growth of the sector has kept pressure upon the capital stock, with no evidence of sustained excess capacity.

Again for the telephone industry as a whole, the distribution of investment outlays to construction, as opposed to machinery and equipment, has remained relatively unchanged over the sample period, with about two-thirds going to machinery and equipment and the remainder to construction. Disaggregated equations estimated for these two components separately are encouragingly consistent, showing the same lag pattern for each category and summing to the previously estimated coefficients for total outlays.

Data for Bell Canada, which accounts for close to two-thirds of total investment expenditures of the telephone industry, are available, and we have employed these to estimate an equation for Bell Canada investment expenditures alone. The resulting equation shows the same pattern of behavior, but with a significantly higher incremental capital/output ratio. This result is consistent with the fact that Bell Canada has been growing more rapidly than the sector as a whole and has been required to invest much more heavily in sophisticated and highly capital-intensive switching and trunking equipment than have many of the smaller firms who make up the balance of the telephone industry.

7. Concluding comment

■ The major conclusions of this analysis have been set out section by section, and there is no need to reiterate them. Here perhaps it is sufficient to observe simply that the results derived constitute only a part of a consistent model of the telephone industry, neglecting any description of financial and regulatory decisions or wage and price formation. Nevertheless, the analysis does yield a consistent pattern flowing from demand decisions back through production decisions to employment and capital formation. Perhaps this model structure may provide a starting point for more detailed dynamic models of the telecommunications sector. Certainly the authors are convinced that such econometric industry models provide a fruitful vehicle for future study of industrial organization and particularly of the impacts of regulatory policies or intervention by government agencies.

References

1. ALLEN, R. G. D. *Mathematical Analysis for Economists.* London: MacMillan Co., 1962.
2. ALMON, C. *The American Economy to 1975.* New York: Harper and Row, 1966.
3. ARROW, K. J., CHENERY, H. B., MINHAS, B., and SOLOW, R. M. "Capital Labor Substitution and Economic Efficiency." *The Review of Economics and Statistics,* Vol. 43, No. 3 (August 1961), pp. 225–50.
4. BALESTRA, P. and NERLOVE, M. "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas." *Econometrica,* Vol. 34, No. 3 (July 1966), pp. 585–612.
5. BAUMOL, W. J. and KLEVORICK, A. K. "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science,* Vol. 1, No. 2 (Autumn 1970), pp. 162–90.
6. BECKMAN, M. J. and SATO, R. "Aggregate Production Functions and Types of Technical Progress: A Statistical Analysis." *The American Economic Review,* Vol. 59, No. 1 (March 1969), pp. 88–101.
7. BISCHOFF, C. W. "Elasticities of Substitution, Capital Malleability, and Distributed Lag Investment Functions." Paper given at the San Francisco meeting of the Econometric Society, December 28, 1966.
8. ———. "Hypothesis Testing and the Demand for Capital Goods." *The Review of Economics and Statistics,* Vol. 51, No. 3 (August 1969), pp. 354–68.
9. CHENERY, H. B. "Overcapacity and the Acceleration Principle." *Econometrica,* Vol. 20, No. 1 (January 1952), pp. 1–28.
10. CHOW, G. C. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica,* Vol. 28, No. 3 (July 1960), pp. 591–605.
11. CLARK, P. G. "The Telephone Industry: A Study in Private Investiment," in W. Leontief, ed., *Studies in the Structure of the American Economy,* New York: Oxford Univ. Press, 1953.
12. DELEEUW, F. "The Demand for Capital Goods by Manufacturers: A Study of Industry Investment Decisions." *Econometrica,* Vol. 30, No. 3 (July 1962), pp. 407–23.
13. DEPARTMENT OF NATIONAL REVENUE. *Taxation Statistics.* Ottawa: Queen's Printer, annual.
14. DHRYMES, P. J. "A Comparison of Productivity Behavior in Manufacturing and Service Industries." *The Review of Economics and Statistics,* Vol. 45, No. 1 (February 1963), pp. 64–69.

15. Dobell, A. R. et al. *Telecommunications in Canada: A Statistical Summary.* Study 2(b) (i) of the Telecommission Studies prepared for the Department of Communications, Information Canada, Ottawa, 1971.

16. Dominion Bureau of Statistics. *Telephone Statistics.* DBS 56–203, Ottawa: Queen's Printer, annual.

17. ———. *Prices and Price Indexes.* DBS 62–001, 62–501, Ottawa: Queen's Printer, annual.

18. Eads, G., Nerlove, M., and Raduchel, W. "A Long-Run Cost Function for the Local Service Airline Industry: An Experiment on Non-Linear Estimation." *The Review of Economics and Statistics*, Vol. 51, No. 3 (August 1969), pp. 258–70.

19. Eisner, R. "Capital Expenditures, Profits, and the Acceleration Principle," in *Models of Income Determination*, National Bureau of Economic Research Study in Income and Wealth, No. 28, Princeton, N. J.: Princeton Univ. Press, 1964, pp. 137–76.

20. ———. "A Distributed Lag Investment Function." *Econometrica*, Vol. 28, No. 1 (January 1960), pp. 1–29.

21. ———. "A Permanent Income Theory for Investment." *The American Economic Review*, Vol. 57, No. 3 (June 1967), pp. 363–90.

22. ———, and Nadiri, M. I. "Investment Behavior and Neoclassical Theory." *The Review of Economics and Statistics*, Vol. 50, No. 3 (August 1968), pp. 369–82.

23. ———, and Strotz, R. H. "Determinants of Business Investment," in *Impacts of Monetary Policies*, Research Monograph, Commission on Money and Credit, Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1963.

24. Evans, M. K. "A Study of Industry Investment Decisions." *The Review of Economics and Statistics*, Vol. 49, No. 2 (May 1967), pp. 151–64.

25. Feldstein, M. S. and Foot, D. K. "The Other Half of Gross Investment: Replacement and Modernization Expenditures." Discussion Paper No. 105, Harvard Institute of Economic Research, January 1970.

26. Fisher, F. M. "The Existence of Aggregate Production Function." *Econometrica*, Vol. 37, No. 4 (October 1969), pp. 553–77.

27. Griliches, Z. "Capital Stock in Investment Functions: Some Problems of Concept and Measurement," in C. F. Christ et al., eds, *Measurement in Economics: Studies in Mathematical Economics and Econometrics*, Stanford, Calif.: Stanford Univ. Press, 1963.

28. ———, and Wallace, N. "The Determinants of Investment Revisited." *International Economic Review*, Vol. 6, No. 3 (September 1965), pp. 311–29.

29. Grunfeld, Y. "The Determinants of Corporate Investment," in A. C. Harberger, ed., *The Demand for Durable Goods*, Chicago: Univ. of Chicago Press, 1960, pp. 211–66.

30. Hall, R. E. and Jorgenson, D. W. "Tax Policy and Investment Behavior." *The American Economic Review*, Vol. 57, No. 3 (June 1967), pp. 391–414.

31. Hickman, B. G. *Investment Demand and U. S. Economic Growth.* Washington, D. C.: The Brookings Institution, 1965.

32. Hildreth, C. and Lu, J. *Demand Relations with Autocorrelated Disturbances.* Technical Bulletin 276, Michigan Agricultural Experimental Station, East Lansing, Michigan, 1960.

33. Houthakker, H. S. and Taylor, L. D. *Consumer Demand in the United States.* 2nd ed. Cambridge, Mass.: Harvard Univ. Press, 1970.

34. Jorgenson, D. W. "Capital Theory and Investment Behavior." *The American Economic Review*, Vol. 53, No. 2 (May 1963), pp. 247–59.

35. ———. "Anticipations and Investment Behavior," in J. S. Duesenberry, E. Kuh, G. Fromm, and L. R. Klein, eds., *The Brookings Quarterly Econometric Model of the United States*, Chicago: Rand McNally & Co., 1965, pp. 35–92.

36. ———. "Rational Distributed Lag Functions." *Econometrica*, Vol. 34, No. 1 (January 1966), pp. 135–49.

37. ———, and Seibert, C. D. "A Comparison of Alternative Theories of Corporate Investment Behavior." *The American Economic Review*, Vol. 58, No. 4 (September 1968), pp. 681–712.

38. ———, and Stephenson, J. A. "Investment Behavior in U. S. Manufacturing, 1947–60." *Econometrica*, Vol. 35, No. 1 (April 1967), pp. 169–220.

39. ———. "The Time Structure of Investment Behavior in U. S. Manufacturing, 1947–60." *The Review of Economics and Statistics*, Vol. 49, No. 1 (February 1967), pp. 16–27.
40. ———. "Issues in the Development of the Neoclassical Theory of Investment Behavior." *The Review of Economics and Statistics*, Vol. 51, No. 3 (August 1969), pp. 346–53.
41. KALDOR, N. "A Model of Economic Growth." *Economic Journal*, Vol. 67, No. 1 (December 1957), pp. 591–624.
42. KLEIN, L. R. *A Textbook of Econometrics*. Evanston, Ill.: Row, Peterson & Co., 1953.
43. KOYCK, L. M. *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland Publishing Co., 1954.
44. KUH, E. "The Validity of Cross Sectionally Estimated Behavior Equations in Time Series Applications." *Econometrica*, Vol. 27, No. 2 (April 1959), pp. 197–214.
45. MARSHAK, J. and ANDREWS, W. H. "Random Simultaneous Equations and the Theory of Production." *Econometrica*, Vol. 12, Nos. 3–4 (July-October 1944), pp. 143–205.
46. MEYER, J. R. and GLAUBER, R. R. *Investment Decisions, Economic Forecasting and Public Policy*. Boston: Harvard Business School, 1964.
47. ———, and KUH, E. *The Investment Decision: An Empirical Study*. Cambridge, Mass.: Harvard Univ. Press, 1957.
48. MODIGLIANI, F. and MILLER, M. "The Cost of Capital, Corporation Finance, and the Theory of Investment." *The American Economic Review*, Vol. 48, No. 3 (June 1958), pp. 261–97.
49. ———. "Some Estimates of the Cost of Capital to the Electric Utility Industry, 1954–57." *The American Economic Review*, Vol. 56, No. 3 (June 1966), pp. 333–91.
50. NERLOVE, M. "Returns to Scale in Electricity Supply," in C. F. Christ et al., eds., *Measurement in Economics· Studies in Mathematical Economics and Econometrics*, Stanford, Calif.: Stanford Univ. Press, 1963.
51. ———. *The Estimation and Identification of Cobb-Douglas Production Functions*. Chicago: Rand McNally & Co., 1965.
52. ———, and WALLIS, K. F. "Use of the Durbin-Watson Statistic in Inappropriate Situations." *Econometrica*, Vol. 34, No. 1 (January 1966), pp. 235–38.
53. OLLEY, R. E. Memorandum on Productivity. Bell Canada, April 1969, Exhibit B-241, Rate Hearings before CTC, 1969.
54. ———. "Productivity Gains in a Public Utility—Bell Canada 1952 to 1967." Paper presented at the Annual Meeting of the Canadian Economics Association, Winnepig, 1970.
55. SALTER, W. E. G. *Productivity and Technical Change*. Cambridge, Mass.: Cambridge Univ. Press, 1966.
56. SOLOW, R. M. "Technical Change and the Aggregate Production Function." *The Review of Economics and Statistics*, Vol. 39, No. 3 (August 1957), pp. 312–20.
57. ———. "Heterogeneous Capital and Smooth Production Functions: An Experimental Study." *Econometrica*, Vol. 31, No. 4 (October 1963), pp. 623–45.
58. ———. "Capital, Labor and Income in Manufacturing," in *The Behavior of Income Shares: Selected Theoretical and Empirical Issues*, National Bureau of Economic Research, Studies in Income and Wealth, No. 27, Princeton, N. J.: Princeton Univ. Press, 1964.
59. ———. "On a Family of Lag Distributions." *Econometrica*, Vol. 28, No. 2 (April 1960), pp. 393–406.
60. TAKAYAMA, A. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 59, No. 3 (June 1969), pp. 255–60.
61. TRANS-CANADA TELEPHONE SYSTEM. *Canada's Telephone Industry in Perspective*.

Investment behavior and the production function

Dale W. Jorgenson

Professor of Economics
Harvard University

In the economic theory of investment behavior the form of the optimal production and investment policy depends critically on the form of technology. The purpose of this paper is to select an appropriate description of technology on the basis of empirical evidence for United States manufacturing industries. The evidence is consistent with a production function characterized by elasticity of substitution equal to unity and constant returns to scale. For this description of technology the optimal policy determines an optimal rate of growth of capital and associated capital/output and labor/output ratios for any set of prices of output, labor input, and capital input. The desired level of capital is a perpetually moving target to which actual capital never converges. The corresponding model of investment policy has been employed extensively in econometric studies of investment behavior. Characterization of the form of optimal investment policy makes it possible to resolve the considerable controversy over the interpretation of econometric models of investment.

1. Introduction

■ In the economic theory of investment behavior the investment process is divided into two separate activities. The first is external to the firm and corresponds to the acquisition of investment goods. The second is internal to the firm and corresponds to installation of these goods. If the market for investment goods is competitive, acquisition takes place at prices that are independent of the level of investment. Installation competes with production of marketable output for the internal resources of the firm, including inputs of capital and labor services. The opportunity cost of a higher level of investment is a lower level of output for given levels of capital and labor input. If the opportunity cost rises at an increasing rate, the average cost of investment goods is increasing, even if all markets are competitive.

In the theory of investment the form of the optimal policy depends critically on the character of installation and production possibilities. Technology may be represented by an efficient installation and production possibility frontier. In this representation, marketable output and the acquisition of investment goods are outputs and capital and labor services are inputs. We may specialize this description of technology by making output Q depend on capital input K, labor input L, and, separately, on the acquisition of investment goods I and

---

Dale W. Jorgenson has been Professor of Economics at Harvard University since 1969. He is associate editor of *The Bell Journal*. Until 1969, he was Professor of Economics at the University of California, Berkeley.

capital input K:

$$Q = F(K,L) - G(I,K), \qquad (1)$$

where F is the production function and G is the installation function. This separable representation of production and installation possibilities has been developed by Lucas and Uzawa[1] as a formalization of costs of adjustment for fixed capital, introduced into the theory of investment by Eisner and Strotz.[2]

If the production and installation functions are characterized by constant returns to scale, proportional changes in any feasible combination of marketable output, acquisition of investment goods and capital and labor input are also feasible. The optimal investment policy determines an optimal rate of growth of capital and associated capital/output and labor/output ratios for any set of prices of output, labor input, and capital input.[3] The theory of investment and production may be contrasted with the conventional theory of production under constant returns to scale with no installation costs. In this theory the maximization of profit determines optimal capital/output and labor/output ratios for any set of prices consistent with zero profits.[4] The scale of output and the optimal level of investment are indeterminate. The introduction of installation costs results in a determinate level of investment and a determinate scale of output for any set of prices. The optimal levels of investment and output at any point of time depend on the amount of capital already accumulated.[5]

Production and installation functions may be characterized by decreasing returns to scale. Treadway has considered a production function with constant or decreasing returns to scale and an installation function depending on the level of acquisition of investment goods alone.[6] As investment increases, the output foregone rises at an increasing rate. The optimal investment policy determines an optimal long-run level of capital input and associated levels of output and labor input for any set of prices of output, labor input, and capital input.[7] The optimal policy converges to the optimal long-run values of output, capital input, and labor input. If returns to scale in production are constant, the level of investment along the optimal path is constant.[8] In the absence of costs of adjustment the optimal policy for production and investment is to move immediately to the optimal long-run values of output, labor input, and capital input.[9]

---

[1] See [45] and [65], respectively.

[2] In [18].

[3] Lucas [45] and Uzawa [65].

[4] In his monograph on Cobb-Douglas production functions, Nerlove states: ". . . Increasing or constant returns to scale are inconsistent with the assumption of *both* perfect competition and profit maximization." This view is, of course, erroneous; constant returns are consistent with perfect competition and profit maximization for any set of prices for which maximum profits are zero, but increasing returns are inconsistent with perfect competition and profit maximization for any set of prices. See Nerlove [55], esp. pp. 8 and 15–16.

[5] For further discussion, see Lucas [45], esp. pp. 325–27.

[6] See [64].

[7] *Ibid.*, esp. pp. 232–37.

[8] Note that Treadway [64] assumes that the installation function depends on the rate of change of capital rather than the level of investment. Using the description of technology in Equation (1) specialized so that installation is independent of capital input, constant returns to scale implies that investment is constant. Capital stock rises (or falls) until investment is equal to replacement.

Under constant returns to scale the responsiveness of the rate of growth of capital and the optimal labor/output and capital/output ratios to changes in the prices of output, labor input, and capital input depends critically on the form of the production function. Similarly, under decreasing returns to scale the responsiveness of the optimal long-run levels of output, capital input, and labor input to price changes depends on the form of the production function. A convenient framework for consideration of responsiveness to price changes is provided by the constant elasticity of substitution (CES) production function proposed by Arrow, Chenery, Minhas and Solow.[10] As a limiting case the CES production function includes the "fixed coefficients" production function of Walras and Leontief.[11] The "fixed coefficients" production function is characterized by labor/output and capital/output ratios that are independent of prices, and has elasticity of substitution zero. Another limiting case of the CES production function is the Cobb-Douglas function[12] with elasticity of substitution equal to unity.

Our first objective in Section 2 is to examine the available evidence on the form of the production function in order to select an appropriate form for the description of technology. A survey article by Nerlove[13] cites more than 40 papers reporting empirical estimates of the elasticity of substitution. From this evidence Nerlove concludes:

> The major finding of this survey is the diversity of results: Even slight variations in the period or concepts tend to produce drastically different estimates of the elasticity. While there seems to be little rhyme or reason for most of these differences, a number of possible sources of bias exist and may account for at least some of the discrepancies.[14]

Subsequent research has provided explanations for many of the discrepancies reported by Nerlove. Griliches and Zarembka have presented evidence that the elasticity of substitution can be estimated reliably from cross sections and from successive cross sections.[15] The estimated elasticity of substitution is not significantly different from unity, suggesting that the most appropriate form of the production function for the description of technology is the Cobb-Douglas form.[16]

The form of the optimal investment policy depends critically on the character of returns to scale. Under constant returns the rate of growth of capital is independent of the size of the firm. Under decreasing returns the rate of growth of capital decreases as the size of the firm increases. The character of returns to scale may be determined by examining the relationship between rate of growth of capital and size of firm. Evidence on returns to scale can also be obtained from the production function. The CES form of the production function has been extended by Dhrymes and Kmenta to incorporate decreasing, constant, or increasing returns to scale.[17] We examine empirical evidence on returns to scale in Section 3. The available evidence from rate of growth of capital and size of firm and from estimation of CES production functions suggests than the most appropriate characterization of returns to scale is that they are constant.

---

[9] See Arrow [2].

[10] In Arrow et al. [3].

[11] See [66] and [44], respectively.

[12] See Douglas [13] and Wicksell [67].

[13] See [56].

[14] *Ibid.*, p. 58.

[15] See Griliches [24, 26, 27], Zarembka [68], and Zarembka and Chernicoff [69].

[16] See Section 2 *infra*.

The optimal investment policy for both constant and decreasing returns is associated with a distributed lag function for net investment.[18] In the absence of gestation lags the distributed lag investment function reduces to the flexible accelerator

$$K_t - K_{t-1} = \lambda[K_t^+ - K_{t-1}], \tag{2}$$

where K_t is the actual level of capital and K_t^+ is the desired level. The interpretation of the desired level of capital in a distributed lag investment function is very different under constant returns to scale and under decreasing returns. Under decreasing returns the desired level of capital is the long-run equilibrium level and depends on the prices of output, labor input, and capital input. Under constant returns there is no long-run equilibrium level of capital; desired capital is a perpetually moving target to which actual capital never converges, provided that prices remain constant.[19] The available evidence suggests that desired capital should be interpreted as a moving target rather than the long-run equilibrium value of capital.[20]

Failure to distinguish between the two alternative interpretations of desired capital, corresponding to constant and decreasing returns, has been an important source of confusion in the theory of investment. In an econometric study of investment behavior that preceded the development of the Lucas-Uzawa model, we gave the following description of optimal production and investment policy:

> . . . We assume that output and employment on the one hand and capital stock on the other are determined by a kind of iterative process. In each period, production and employment are set at levels given by the . . . marginal productivity condition [for labor] and the production function with capital stock fixed at its current level; demand for capital is set at the level given by the . . . marginal productivity condition [for capital], given output and employment.[21]

The model was completed by a distributed lag function for net investment. The complete model corresponds precisely to optimal production and investment policy under constant returns in production and installation. Subsequently, we presented a derivation of the marginal productivity conditions for labor and capital based on the conventional theory of production with no explicit installation costs and decreasing returns.[22] Installation costs were introduced implicitly through the distributed lag function for investment. The resulting econometric model is based on the iterative process described above. A derivation of this model incorporating installation costs explicitly with constant returns to scale in both production and installation is obviously much more satisfactory than the original deviation. We discuss the form of the optimal investment policy in greater detail in Section 4.

---

[17] See [11] and [42], respectively.

[18] For approximation under diminishing returns, see Lucas [46].

[19] Desired capital is proportional to the level of output; under constant returns to scale output and capital grow at the same rate.

[20] See Section 3 *infra*.

[21] Jorgenson [30], p. 249.

[22] In Jorgenson [29], esp. pp. 45 and 53.

2. The form of the production function

■ In the absence of installation costs the CES production function may be written

$$Q = \gamma[\delta K^{-\rho} + (1 - \delta)L^{-\rho}]^{-\nu/\rho}, \tag{3}$$

where γ is the *efficiency* parameter, δ the *distribution* parameter, ρ the *substitution* parameter, and ν the *scale* parameter. The elasticity of substitution σ is defined as the elasticity of the capital/labor ratio with respect to the marginal rate of substitution between capital and labor,[23]

$$\sigma = \frac{1}{1 + \rho}. \tag{4}$$

The scale parameter is, of course, the degree of homogeneity of the production function.[24]

If the substitution parameter ρ is assigned the value of minus one, the factors of production are perfect substitutes; the limiting production function as the substitution parameter ρ becomes infinite is the fixed coefficients production function; the limiting function as the substitution parameter ρ goes to zero is the Cobb-Douglas production function.[25] For these three forms of the production function the elasticity of substitution σ is infinity, zero, and unity, respectively. All three forms of the production are, of course, homogeneous of degree ν.

The CES production function may be approximated by an expression quadratic in the logarithms of the variables. To derive this approximation we write the function in the form

$$\ln Q = \ln\gamma - \frac{\nu}{\rho}\ln[\delta e^{-\rho \ln K} + (1 - \delta)e^{-\rho \ln L}].$$

Expanding this function in Taylor's series around $\ln Q = \ln K = \ln L = 0$, we obtain

$$\ln Q = \ln\gamma + \nu\delta \ln K + \nu(1 - \delta)\ln L - \rho\nu\delta(1 - \delta)(\ln K - \ln L)^2. \tag{5}$$

This approximation is identical to the approximation linear in the substitution parameter ρ derived by Kmenta.[26] The first three terms correspond to a Cobb-Douglas production function. The last term represents the contribution of the substitution parameter ρ; if this parameter is zero, the approximation reduces to the Cobb-Douglas function.

Under constant returns to scale the CES production function reduces to

$$Q = \gamma[\delta K^{-\rho} + (1 - \delta)L^{-\rho}]^{-1/\rho}, \tag{6}$$

---

[23] See Allen [1], pp. 340–45.

[24] $\lambda^{\nu}Q = \gamma[\delta(\lambda K)^{-\rho} + (1 - \delta)(\lambda L)^{-\rho}]^{-\nu/\rho}$.

[25] These arguments are due to Arrow et al. [3], pp. 230–31.

[26] In [42], pp. 180–81.

the form originally proposed by Arrow et al.[27] The marginal conditions for equilibrium with no costs of adjustment are

$$
\begin{aligned}
\frac{\partial Q}{\partial L} &= \gamma^{-\rho}(1-\delta)\left(\frac{Q}{L}\right)^{1+\rho} = \frac{q_L}{q_Q}, \\
\frac{\partial Q}{\partial K} &= \gamma^{-\rho}\delta\left(\frac{Q}{K}\right)^{1+\rho} = \frac{q_K}{q_Q},
\end{aligned}
\tag{7}
$$

where q_Q is the price of output, q_K the price of capital input, and q_L the price of labor input. With costs of adjustment, the second marginal condition is replaced by

$$
\frac{\partial Q}{\partial K} = \frac{q_K}{q_Q} + (r+\delta)\frac{\partial G}{\partial I}, \tag{8}
$$

where r is the rate of interest and δ the rate of replacement of investment goods.[28] In this form the marginal condition may be approximated by a distributed lag investment function.[29]

We turn next to an examination of the empirical evidence on the form of the production function. The empirical literature through 1967 has been surveyed by Griliches, Lucas, and Nerlove.[30] Griliches' general conclusion is that

> in reviewing previous estimates of the elasticity of substitution in U. S. manufacturing industries we are faced with two conflicting sets of estimates. The studies based on cross-sectional data yield estimates which are on the whole not significantly different from unity. The time series studies report, on the average, substantially lower estimates.[31]

In comparing the two sets of estimates Lucas observes that the time series results may be biased by misspecification of the lag structure, while the cross section results may be biased by omission of regional differentials in the price of output and quality of labor.[32] Griliches and Zarembka have attempted to remove biases from both of the data sources.[33] Bischoff has attempted to estimate the elasticity of substitution in a distributed lag investment function with careful attention to correct specification of the lag structure.[34] We consider the evidence presented by Bischoff, Griliches, and Zarembka, together with earlier estimates surveyed by Griliches, Lucas, and Nerlove. We first consider evidence based on a maintained hypothesis of constant returns to scale.

The major conclusions of the original empirical study by Arrow et al. are that (1) the elasticity of substitution is significantly different from zero, and (2) the elasticity of substitution is significantly below unity. The first of these conclusions holds for every industry group included in the original study. The second holds for only ten of

---

[27] See [3], pp. 230.

[28] See Lucas [45], formula (9), p. 325.

[29] See note 18 above.

[30] See Griliches [27], pp. 285–90, Lucas [47], pp. 232–41 and 245–47, and Nerlove [56] *passim*.

[31] [27], p. 285.

[32] In Lucas [47], pp. 236–41 and 261–65.

[33] See Griliches [27], pp. 290–97 and Zarembka [68], pp. 48–49.

[34] In [8].

TABLE 1

COMPARISON OF ESTIMATES OF THE ELASTICITY OF SUBSTITUTION FOR MANUFACTURING INDUSTRIES BY ACMS AND BY FUCHS FROM INTERNATIONAL CROSS SECTIONS, 1950–1955

INDUSTRY GROUP	(1) ACMS	(2) FUCHS
FOOD		
DAIRY	0.72 (0.07)	0.90 (0.08)
FRUIT AND VEGETABLE CANNING	0.86 (0.08)	1.09 (0.10)
GRAIN AND MILL PRODUCTS	0.91 (0.10)	1.32 (0.17)
BAKERY PRODUCTS	0.90 (0.07)	1.07 (0.11)
SUGAR	0.78 (0.12)	0.90 (0.18)
TOBACCO	0.75 (0.15)	1.22 (0.21)
TEXTILE		
SPINNING AND WEAVING	0.81 (0.07)	0.98 (0.10)
KNITTING MILLS	0.79 (0.06)	0.95 (0.08)
LUMBER	0.86 (0.07)	1.08 (0.14)
FURNITURE	0.89 (0.04)	1.04 (0.09)
PAPER	0.97 (0.10)	0.91 (0.18)
PRINTING AND PUBLISHING	0.87 (0.06)	1.02 (0.09)
CHEMICALS		
BASIC CHEMICALS	0.83 (0.07)	1.11 (0.10)
MISCELLANEOUS CHEMICALS	0.90 (0.06)	1.06 (0.09)
FATS AND OILS	0.84 (0.09)	1.06 (0.18)
LEATHER	0.86 (0.06)	0.98 (0.10)
STONE, CLAY, AND GLASS		
CLAY	0.92 (0.10)	0.66 (0.20)
GLASS	1.00 (0.08)	1.27 (0.10)
CERAMICS	0.90 (0.04)	1.08 (0.13)
CEMENT	0.92 (0.15)	1.31 (0.22)
PRIMARY METALS		
IRON AND STEEL	0.81 (0.05)	0.76 (0.11)
NONFERROUS METALS	1.01 (0.12)	0.94 (0.20)
FABRICATED METALS	0.90 (0.09)	1.01 (0.17)
ELECTRICAL MACHINERY	0.87 (0.12)	1.03 (0.21)

SOURCES: (1) ARROW ET AL. [3], TABLE 2, P. 227,
(2) FUCHS [20], TABLE 2, P. 438.

TABLE 2

ESTIMATES OF THE ELASTICITY OF SUBSTITUTION FOR TWO–DIGIT MANUFACTURING INDUSTRIES IN THE U S BY GRILICHES

INDUSTRY GROUP	(1) CROSS SECTION	(2) LAGGED ADJUSTMENT	(3) SERIAL CORRELATION
FOOD	0.91 (0.10)	1.01	*
TEXTILE	0.94 (0.17)	1.11	1.09 (0.54)
APPAREL	1.06 (0.19)	0.84	0.63 (0.21)
LUMBER	1.07 (0.06)	1.11	1.18 (0.20)
FURNITURE	1.04 (0.07)	0.99	*
PAPER	1.67 (0.30)	1.30	*
PRINTING	0.83 (0.18)	0.45	0.68 (0.28)
CHEMICALS	0.71 (0.22)	0.59	0.70 (0.40)
PETROLEUM	**	**	**
RUBBER	1.28 (0.42)	2.21	0.90 (0.43)
LEATHER	0.84 (0.26)	1.60	1.16 (0.40)
STONE, CLAY, GLASS	0.91 (0.19)	0.77	1.88 (0.49)
PRIMARY METALS	1.41 (0.42)	3.49	2.37 (0.47)
FABRICATED METALS	0.85 (0.14)	1.17	1.20 (0.28)
MACHINERY, EXCL. ELECTRICAL	1.24 (0.38)	2.40	2.00 (0.29)
ELECTRICAL MACHINERY	0.66 (0.31)	0.40	1.53 (0.53)
TRANSPORTATION EQUIPMENT	0.96 (0.55)	1.09	*
INSTRUMENTS	0.75 (0.43)	0.82	*

*CONTRADICTS SERIAL CORRELATION MODEL; SEE GRILICHES [27], TABLE 3, NOTE C, P. 294

** NO SIGNIFICANT RELATIONSHIPS FOUND.

SOURCE: (1) – (3) GRILICHES [27], TABLE 3, PP. 293-94

24 industry groups.[35] Fuchs has run regressions for the same groups, inserting a dummy variable to allow for differences in level of development among the countries included in the sample.[36] This variable helps to remove biases due to variations in labor quality and the price of output between developed and less developed countries. The results confirm the first conclusion of Arrow et al. but not the second. For only two industry groups is the estimated elasticity of substitution significantly different from unity, and in one of these cases it is significantly *above* unity. The two sets of results are compared in Table 1. We conclude that the finding of Arrow et al. that the elasticity of substitution is less than unity results from a bias due to variations in labor quality or the price of output between developed and less developed countries. When this bias is removed, the estimated elasticity of substitution is not significantly different from unity.

---

[35] See Arrow et al. [3], Table 2, p. 227.

[36] See [20].

The study of Arrow et al. is based on international cross sections for three-digit industries within manufacturing. We next review evidence on the elasticity of substitution for two-digit manufacturing industries in the U. S. We begin by considering evidence based on the marginal productivity condition for labor alone reported by Griliches and Zarembka for cross sections of states in 1957 and 1958.[37] Griliches first estimates the elasticity of substitution from a cross section of states for 1958. The results are given in Column (1) of Table 2.[38] Griliches fits alternative dynamic models to successive cross sections of states for 1957 and 1958, one a lagged adjustment model and the other based on first-order serial correlation in the error term. The lagged adjustment model includes the lagged value of the dependent variable as one of the regressors. The serial correlation model includes lagged values of both dependent and independent variables as regressors. The results are given in Columns (2) and (3) of Table 2.[39] Griliches summarizes these results as follows:

> The second set of σ estimates is based on the partial adjustment equation, while the third is based on the serial correlation model. In 12 out of seventeen cases the latter model is the one consistent with the data. In general, all of the estimated σ's are not very (statistically) different from unity, the significant deviations if anything occurring above unity rather than below it.[40]

Serial correlation may be explained as the result of persistence of omitted variables such as regional differentials in the price of output or the quality of labor. Griliches fits the same models to data for all manufacturing industries taken together, concluding that the serial correlation model is more appropriate than the lagged adjustment model. The estimated elasticity of substitution is above unity.[41]

Zarembka has estimated the elasticity of substitution for two separate cross sections of states for 1957 and 1958, taking into account possible serial correlation in the errors for the two periods. Data on labor input and wage rates are corrected for quality change, but the effects of differences in product prices or regional differences in efficiency appear as a possible source of serial correlation.[42] The statistical model is similar to that of Griliches, except that Griliches maintains the hypothesis of equality of the elasticities of substitution for the two time periods, while Zarembka tests this hypothesis. The results provide a direct empirical test of one of the conclusions reached by Nerlove, namely, that changes in time period result in substantial changes in the elasticity of substitution.[43] The estimated elasticities for the two periods and an *F*-ratio for a test that the parameters are the same are given in Columns (1), (2), and (3) of Table 3. The estimated elasticities for the two periods are not significantly different for 11 of the 13 industries at a significance level of 0.025. The obvious difficulty with Nerlove's comparison of results[44] is that variances and covariances of estimators for different time periods are ignored.

---

[37] See Griliches [27], pp. 290–94, and Zarembka [68] *passim.*

[38] Griliches [27], Table 3, p. 293.

[39] *Ibid.*, Table 3, p. 294.

[40] *Ibid.*, p. 292.

[41] *Ibid.*, pp. 294–97.

[42] See [68], p. 49.

[43] See [56], p. 59.

[44] *Ibid.*, pp. 68–70.

TABLE 3

ESTIMATES OF THE ELASTICITY OF SUBTITUTION FOR TWO-DIGIT MANUFACTURING INDUSTRIES IN THE U.S. BY ZAREMBKA

INDUSTRY GROUP	(1) 1957	(2) 1958	(3) F-RATIO	(4) BOTH YEARS
FOOD	0.96 (0.19)	1.01 (0.14)	0.12	1.01 (0.14)
TEXTILE	0.79 (0.17)	0.91 (0.18)	0.73	0.84 (0.16)
APPAREL	0.96 (0.21)	0.86 (0.18)	0.87	0.88 (0.19)
FURNITURE	1.13 (0.19)	1.07 (0.12)	0.11	1.08 (0.11)
PAPER	0.88 (0.34)	0.72 (0.25)	0.40	0.77 (0.26)
CHEMICALS	1.15 (0.24)	0.92 (0.27)	2.88	1.11 (0.24)
PETROLEUM	1.06 (0.40)	0.31 (0.58)	3.18	1.00 (0.42)
STONE, CLAY, GLASS	0.77 (0.25)	0.64 (0.23)	0.48	0.70 (0.22)
PRIMARY METALS	1.17 (0.26)	1.81 (0.26)	12.26	1.43 (0.25)
FABRICATED METALS	0.59 (0.14)	0.67 (0.15)	0.35	0.63 (0.13)
MACHINERY, EXCL ELECTRICAL	1.36 (0.32)	1.13 (0.28)	0.93	1.22 (0.26)
ELECTRICAL MACHINERY	0.63 (0.24)	0.21 (0.21)	5.89	0.38 (0.21)
TRANSPORTATION EQUIPMENT	1.49 (0.49)	1.47 (0.33)	0.01	1.46 (0.32)

SOURCE: (1)-(4) ZAREMBKA [68], TABLE 1, P. 50

Proceeding conditionally on the hypothesis that elasticities of substitution are the same for the two years 1957 and 1958, Zarembka estimates the elasticity from data for both years, again taking into account the possibility of serial correlation in errors for the two years. The statistical model is identical to that underlying Griliches' serial correlation model, but Zarembka's estimates are more efficient. Griliches ignores constraints on the parameters of the serial correlation model and estimates this model by ordinary least squares.[45] Zarembka's results are presented in Column (4) of Table 3. For only two of the industry groups is the estimated elasticity of substitution significantly different from unity at a significance level of 0.025. The test statistic for this hypothesis and the *F*-ratio of Column (3) are distributed independently since the two hypotheses are "nested."[46] The overall level of significance for the two tests is, of course, 0.05. Zarembka's results support Griliches' conclusion that the elasticity of substitution is equal to unity. For all but one industry group the estimated elasticity is significantly different from zero, corroborating the results of Arrow et al. Zarembka concludes

> . . . that the elasticity of substitution does not in general depart significantly from unity. Therefore, the more complicated CES production function simplifies to the Cobb-Douglas.[47]

The results of Griliches and Zarembka are for data on output and labor input for successive cross sections of states in 1957 and 1958.

---

[45] See [27], pp. 291–94.

[46] See Scheffé [58], esp. pp. 44–45 and 111–12.

[47] [68], p. 51; see also Zarembka and Chernicoff [69], and Griliches [26], p. 151, note 1.

TABLE 4

ESTIMATES OF THE ELASTICITY OF SUBSTITUTION FOR TWO-DIGIT MANUFACTURING INDUSTRIES IN THE U S BY BELL, DHRYMES, MINASIAN, AND SOLOW

INDUSTRY GROUP	(1) BELL	(2) DHRYMES	(3) MINASIAN	(4) SOLOW
FOOD	0.65 (0.12)	0.82 (0.16)	0.58 (0.16)	0.69 (0.22)
TOBACCO			3.46 (0.52)	1.96 (0.30)
TEXTILE	0.98 (0.19)	0.97 (0.15)	1.58 (0.35)	1.27 (0.15)
APPAREL	1.37 (0.36)	1.21 (0.20)		1.01 (0.13)
LUMBER	0.94 (0.05)	0.88 (0.07)	0.94 (0.11)	0.99 (0.09)
FURNITURE	1.03 (0.11)	1.17 (0.13)	1.09 (0.23)	1.12 (0.11)
PAPER	1.53 (0.31)	1.43 (0.45)	1.60 (0.35)	1.77 (1.01)
PRINTING	1.12 (0.22)	1.00 (0.21)		1.02 (0.21)
CHEMICALS	0.69 (0.25)	0.87 (0.27)		0.14 (0.95)
PETROLEUM	1.32 (0.43)	0.89 (0.55)	–0.54 (1.06)	1.45 (0.71)
RUBBER	1.35 (0.31)	1.56 (0.35)	0.82 (0.29)	1.48 (0.88)
LEATHER	0.90 (0.25)	0.86 (0.26)	0.96 (0.29)	0.89 (0.27)
STONE, CLAY, GLASS	0.86 (0.18)	1.03 (0.19)	0.59 (0.25)	0.32 (0.46)
PRIMARY METALS	1.07 (0.25)	0.77 (0.40)	0.92 (0.24)	1.87 (0.25)
FABRICATED METALS	0.80 (0.21)	0.56 (0.20)		0.80 (0.30)
MACHINERY, EXCL. ELECTRICAL	0.54 (0.27)	0.75 (0.47)	0.31 (0.21)	0.63 (0.45)
ELECTRICAL MACHINERY	0.78 (0.23)	0.59 (0.36)	1.26 (0.33)	0.37 (0.54)
TRANSPORTATION EQUIPMENT	1.89 (0.50)	1.24 (0.64)	2.04 (0.49)	0.06 (0.82)
INSTRUMENTS	0.74 (0.33)			1.59 (0.15)

SOURCES (1) BELL [7], TABLE 1, P. 329,
(2) DHRYMES AND ZAREMBKA [12], TABLE 1, COLUMN "σ_1," P. 116,
(3) MINIASIAN [52], TABLE 1, COLUMN "η_{w_1}," P. 267,
(4) SOLOW [63], TABLE 2, COLUMNS (2) AND (3), P. 113.

These data have also been analyzed by Bell and Minasian for 1957 and by Dhrymes for 1958.[48] Solow has estimated the elasticity of substitution for a cross section of regions for 1956.[49] Bell's results are given in Column (1) of Table 4. For 18 industry groups the estimated elasticity of substitution is significantly different from unity for only one industry. The results originally reported by Dhrymes are described as based on a regression of the output/labor ratio on the wage rate.[50] Subsequently, Dhrymes and Zarembka discovered that the reported results are for a regression of the wage rate on the output/labor ratio.[51] Estimates for 1958 based on a correction of this error are given in Column (2) of Table 4. The results are comparable to those of Bell; only one of the estimated elasticities of substitution is significantly different from unity. Zarembka's tests for differences in elasticities of substitution for the two years, given in Table 3, are based on data similar to those of Bell and Dhrymes.

---

[48] See Bell [7], Minasian [52], and Dhrymes [11].

[49] See [63].

[50] See [11], pp. 357–58.

[51] See [12].

Minasian has estimated the elasticity of substitution from data on output and labor input for a cross section of states for 1958; the results are given in Column (3) of Table 4.[52] Nerlove has compared these results with those of Dhrymes as follows:

> . . . Results obtained by Dhyrmes . . . are based on logarithmic regressions of value added per unit of labor on the wage rate across states in 1957. Even though these results appear to be based on substantially the same data (individual states, 1957) as those obtained by Minasian, it is clear . . . that they differ very substantially from the corresponding estimates obtained by Minasian. For example, Minasian's estimate of σ for pulp, paper, and products is nearly eight times Dhrymes', while his estimate of σ for primary metal products is nearly ten times Dhrymes' Conversely, Minasian's estimate for rubber products is double Dhrymes'. The only explanation for these gross differences appears to be slight variation in the basic series employed.[53]

The correct explanation is more straightforward, namely that Dhrymes' reported results are erroneous. For the corrected results a comparison shows that Minasian's estimated σ for pulp, paper, and products is 1.60 with a standard error of 0.35. Dhrymes' estimated σ is 1.43 with a standard error of 0.45. Minasian's result for primary metal products is 0.92 with a standard error of 0.24; Dhrymes' estimate is 0.77 with a standard error of 0.40. Dhrymes' corrected estimate for rubber products of 1.56 exceeds Minasian's 0.82.

Solow's estimates of the elasticity of substitution, based on a cross section of regions for 1956, are given in Column (4) of Table 4.[54] For a total of 18 industry groups Minasian finds two estimates significantly below unity, three significantly above, and 13 not significantly different from unity. Solow finds no estimates significantly below unity, one significantly above, and 17 not significantly different from unity. Griliches has compared his results for a cross section of states for 1958 with those of Minasian and Solow as follows:

> The first set of σ estimates [given in Table 2, Column (1)], is comparable to, though substantially better than (in terms of fit and t-ratios), the Minasian and Solow estimates and is generally of the same order of magnitude. Only one of these σ's (out of 17) is significantly different from unity, and that one is above unity.[55]

Griliches has also compared his results with the original results reported by Dhrymes; in discussing the discrepancy, Griliches reports that

> the major difference . . . is in the definition of labor and the associated wage rate. Dhrymes defines labor as total employment and the associated "wage rate" as total payrolls/total employment. I define the "wage rate" as the average wage rate per hour of production workers (given by production worker wages/total man hours of production workers) and the associated "labor input" measure as total man hours in production worker hour equivalents. . . .[56]

---

[52] Minasian [52], p. 267.
[53] [56], p. 74.
[54] See Solow [63], p. 113.
[55] [27], p. 292.
[56] [24], p. 608.

This difference is adduced as an explanation of the discrepancy between Griliches' results (for states in 1958) and Dhrymes' results (for states in 1957). Dhrymes' corrected results are closely comparable to those of Griliches. Neither the difference in concept nor the difference in time period results in substantial discrepancies between the two sets of results.

We conclude that evidence on the elasticity of substitution for two-digit manufacturing industries in the United States from data on output and labor input is consistent with the Cobb-Douglas form of the production function. Estimated elasticities of substitution for cross sections of states in 1957 and 1958 by Griliches and Zarembka, taking into account possible serial correlation in the residuals for the two years, are not significantly different from unity. These results are consistent with those of Bell, Dhrymes, and Minasian for cross sections of states for a given year and with those of Solow for a cross section of regions. Second, this evidence is inconsistent with the "fixed coefficients" production function, that is, with constant capital/output and labor/output ratios. The empirical results support the conclusion of Arrow et al. that the elasticity of substitution is significantly different from zero. Nerlove's conclusion, cited above, that slight variations in the time period or the concepts of measurement tend to produce "drastically" different estimates of the elasticity of substitution is based on a failure to use appropriate tests of significance in comparing results of different time periods and on Dhrymes' error in reporting his results.[57]

We have reviewed evidence on the elasticity of substitution from data on output and labor input for successive cross sections. Analogous relationships have been fitted to data on capital input. We turn now to a review of this evidence, focusing on the results of Bell and Dhrymes.[58] Bell's estimates are for cross sections of states in 1958; the estimates are based on a regression of the logarithm of the ratio of the share of labor to the share of capital on the logarithm of the capital/labor ratio. In effect, Bell treats the price variable as dependent, while Griliches and Zarembka treat the quantity variable as dependent. Bell's results are given in Column (1) of Table 5. Dhrymes' estimates are for cross sections of states in 1957; Dhrymes' original results are described as based on a regression of the ratio of output to capital on the rental rate for capital, but Zarembka subsequently discovered that the reported results were for a regression of the rental rate on the output/capital ratio. The original results are given in Column (2) of Table 5. Results for a regression of the output/capital ratio on the rental value are presented in Column (3) of Table 5.

A comparison of Bell's estimates given in Column (1) of Table 5 with Dhrymes' estimates given in Column (2) shows that the estimated elasticity of substitution exceeds that from corresponding data on output and labor input and estimates exceeding unity predominate. Reversing the direction of the regression for Dhrymes' data produces the results given in Column (3); for these estimates price is treated as an independent variable, as in the estimates of Griliches and Zarembka examined above. The estimated elasticities

---

[57] See Nerlove [56], p. 58.

[58] See [6] and [11], respectively.

TABLE 5

ESTIMATES OF THE ELASTICITY OF SUBSTITUTION FOR TWO–DIGIT MANUFACTURING INDUSTRIES IN THE U. S. FROM DATA ON CAPITAL BY BELL AND DHRYMES

INDUSTRY GROUP	(1) BELL	(2) DHRYMES ORIGINAL	(3) DHRYMES CORRECTED
FOOD	0.91	0.97 (0.13)	0.77 (0.07)
TEXTILE	1.16	1.03 (0.15)	0.81 (0.10)
APPAREL	1.54	1.03 (0.18)	0.77 (0.08)
LUMBER	1.01	1.10 (0.11)	0.64 (0.08)
FURNITURE	1.15	1.39 (0.06)	0.74 (0.03)
PAPER	1.56	0.64 (0.08)	1.16 (0.11)
PRINTING	1.11	1.11 (0.06)	0.83 (0.06)
CHEMICALS	1.05	1.03 (0.06)	0.91 (0.05)
PETROLEUM	1.71	1.31 (0.08)	0.77 (0.04)
RUBBER	1.35	1.04 (0.14)	0.81 (0.10)
LEATHER	1.09	1.13 (0.12)	0.82 (0.08)
STONE, CLAY, GLASS	1.24	0.89 (0.08)	0.97 (0.06)
PRIMARY METALS	1.28	0.97 (0.14)	0.84 (0.10)
FABRICATED METALS	1.34	0.95 (0.15)	0.72 (0.10)
MACHINERY, EXCL. ELECTRICAL	1.57	0.24 (0.70)	0.36 (0.12)
ELECTRICAL MACHINERY	1.34	0.62 (0.35)	0.73 (0.06)
TRANSPORTATION EQUIPMENT	1.62	0.91 (0.10)	0.69 (0.13)
INSTRUMENTS	1.50		

SOURCES (1) BELL [6], TABLE 1, FIRST COLUMN " b," P. 329;
(2) DHRYMES [11], TABLE 1, COLUMN "σ_2," P. 358,
(3) DHRYMES AND ZAREMBKA [12], TABLE 1, COLUMN "σ_2," P. 116

are predominantly lower than the elasticities estimated from data on output and labor input. This evidence overturns another of Nerlove's conclusions, namely, that

> . . . elasticities of substitution based on logarithmic regressions of value added per unit of capital stock (essentially book value) on the rate of return to capital (computed as a residual) . . . are uniformly higher than those based on the more usual estimation procedure.[59]

One possible explanation of differences between estimates based on labor data and estimates based on capital data is easily discarded. The rental price of capital is measured by the ratio of property income to the quantity of capital. Errors in measurement of the quantity of capital results in corresponding biases in the measured output/capital ratio and the measured rental rate; biases of this type would bias the estimated elasticity of substitution toward unity; only five of the 17 estimates from capital data are between unity and the corresponding estimates from labor data, which are unaffected by the bias due to errors in the measurement of capital.[60] Discrepancies between labor and capital estimates of the elasticity of substitution cannot be explained by errors in the measurement of capital.

---

[59] [56], p. 74. Eisner and Nadiri have employed this conclusion and the conclusions cited above, notes 14 and 53, to support their estimate of the elasticity of substitution from time series. See Eisner and Nadiri [17], esp. p. 217. See also Eisner [15], esp. p. 747, and Coen [10], esp. p. 372. None of Nerlove's conclusions stand up to examination of the empirical evidence; accordingly, the arguments by Coen, Eisner, and Nadiri based on these conclusions are not supported by the evidence. For further dicsussion, see Section 3 *infra.*

An alternative explanation is that marginal conditions with no costs of adjustment are appropriate for labor but not for capital.[61] The marginal condition for capital may be written

$$\gamma^{-\rho}\delta\left(\frac{Q}{K}\right)^{1+\rho} = \frac{q_K}{q_Q}\left(1 + \frac{q_Q}{q_I}\frac{\partial G}{\partial I} + \frac{q_Q}{q_K}\frac{\partial G}{\partial K}\right) = \frac{q_K}{q_Q}A\,,$$

where q_I is the price of investment goods and

$$A = \left(1 + \frac{q_Q}{q_I}\frac{\partial G}{\partial I} + \frac{q_Q}{q_K}\frac{\partial G}{\partial K}\right)$$

represents a wedge between the ratio of the price of capital services to the price of output and the marginal product of capital due to costs of adjustment. Adding an error to this equation, say ϵ_K, the ratio of output to capital services may be expressed in the form

$$\ln\frac{Q}{K} = \sigma[\rho\ln\gamma - \ln\delta] + \sigma\left[\ln\frac{q_K}{q_Q} + \ln A\right] + \epsilon_K\,.$$

We assume that the variable representing the effect of costs of adjustment, $\ln A$, is distributed independently of the price variable, $\ln\frac{q_K}{q_Q}$, and the error in the equation, ϵ_K. Under this assumption the appropriate independent variable in a regression function for the quantity variable, $\ln\frac{Q}{K}$, is the sum of the price variable and the effect of costs of adjustment. Use of the price variable as the independent variable involves an error of measurement and results in a bias toward zero in the estimated elasticity of substitution. This explains the difference between the estimated elasticity of substitution from capital data given by Dhrymes and Zarembka in Column (3) of Table 5 and the corresponding estimate from labor data in Column (2) of Table 4.

Reversing the relationship determining the output/capital ratio we obtain

$$\ln\frac{q_K}{q_Q} = -\rho\ln\gamma + \ln\delta + \frac{1}{\sigma}\ln\frac{Q}{K} - \ln A - \frac{1}{\sigma}\epsilon_K\,.$$

For a positive elasticity of substitution, $\sigma > 0$, the error in this equation, $-\ln A - \frac{1}{\sigma}\epsilon_K$, is negatively correlated with the right-hand side variable, $\ln\frac{Q}{K}$, so that the estimated value of the coefficient of this variable, $\frac{1}{\sigma}$, is biased downward and the estimated value of the elasticity of substitution σ is biased upward. This explains the difference between the estimated elasticity of substitution originally given by Dhrymes and presented in Column (2) of Table 5 and the estimate from labor data by Dhrymes and Zarembka in Table 4.

---

[60] Griliches [27], p. 287. note 28, presents an analysis of bias for errors in the measurement of labor, the analysis for capital is strictly analogous.

[61] See formula (8) above.

Turning to the relationship fitted by Bell, the marginal productivity condition for labor may be written

$$\ln\frac{Q}{L} = \sigma[\rho \ln\gamma - \ln(1 - \delta)] + \sigma \ln\frac{q_L}{q_Q} + \epsilon_L ,$$

where ϵ_L is the error in the equation. Combining this condition with the corresponding condition for capital and solving for the logarithm of the ratio of the share of labor to the share of capital, we obtain

$$\ln\frac{q_L L}{q_K K} = -\ln\frac{\delta}{1 - \delta} + \rho\left(\ln\frac{Q}{L} - \ln\frac{Q}{K}\right) + \ln A + \frac{1}{\sigma}(\epsilon_K - \epsilon_L) .$$

For a positive elasticity of substitution, $\sigma > 0$, the error in this equation, $\ln A + \frac{1}{\sigma}(\epsilon_K - \epsilon_L)$, is negatively correlated with the right-hand side variable,

$$\ln\frac{Q}{L} - \ln\frac{Q}{K} = \ln\frac{K}{L} ,$$

so that the estimated value of the coefficient is biased downward and the estimated value of the elasticity of substitution (4) is biased upward.

We conclude that marginal conditions with no costs of adjustment are inappropriate for capital input. On the assumption that the effect of costs of adjustment, $\ln A$, is distributed independently of the price variables, $\ln\frac{q_K}{q_Q}$ and $\ln\frac{q_L}{q_Q}$, and the errors in the equations, ϵ_K and ϵ_L, discrepancies between Dhrymes' results from capital and labor data, may be attributed to biases due to omission of costs of adjustments in the relationship fitted to capital data. This omission results in a downward bias in the estimated elasticity of substitution for Dhrymes' regression of the logarithm of the output/capital ratio on the logarithm of the corresponding price ratio, given in Column (3) of Table 5. It results in an upward bias in the reverse regression, given in Column (2) of Table 5. The omission of costs of adjustment results in an upward bias in the estimated elasticity for Bell's regression of the logarithm of the ratio of the logarithm of the share of labor to the share of capital on the logarithm of the capital/labor ratio, given in Column (1) of Table 5.

3. Returns to scale

■ Up to this point we have presented evidence on the form of the production function from cross sections. This evidence is based on the maintained hypothesis that returns to scale are constant. We turn now to a review of evidence on returns to scale. We first consider evidence that is independent of the form of the production function. Second, we review estimates of the elasticity of substitution from cross sections and time series, not based on a maintained hypothesis of constant returns. The elasticity of substitution and the scale parameter or degree of homogeneity of the production function can be estimated simultaneously.

By far the most careful study of economies of scale for the manufacturing industries is that of Bain. For a sample of 20 manufacturing industries Bain obtained detailed engineering estimates of optimal plant size and compared the results with the average size of plants in the largest size class of plants from the 1947 Census of Manufactures.[62] The results are quite similar and suggest that the optimal plant size is small as a percentage of total industry capacity.[63] Second, returns to scale at the plant level beyond the minimum efficient plant size are constant.[64] Third, there is no evidence of substantial economies of multi-plant operation.[65] The overall conclusion is that returns to scale at the firm level are increasing up to a firm size equal to the minimum optimal size of plant; beyond that point returns to scale are constant. Bain summarizes the findings as follows:

> Three general findings concerning scale curves have emerged from this sort of study. First, in the usual industry, economies of large scale are realized as the firm expands from miniscule size to some identifiable minimal optimal scale, but not beyond that scale. Second, unit costs tend to remain constant at the optimal level as the firm grows still larger, and diseconomies of very large scale ordinarily do not seem to be encountered. Third, most of the scale economies observed are economies of large-scale plant. Although these findings apply specifically to scale curves measuring only production and distribution costs, they are roughly sustained when sales promotion costs are included, except that in some product-differentiated industries smaller firms seem to encounter diseconomies of very large-scale sales promotion.[66]

The form of the optimal policy for capital accumulation depends critically on the character of returns to scale. For constant returns to scale the optimal policy determines labor/output and capital/output ratios and the rate of growth of capital. The optimal policy has the important implications that no optimal scale for the firm exists and that the rate of growth of the firm is independent of size.[67] For decreasing returns to scale the optimal policy determines long-run equilibrium values of output and of labor and capital input so that an optimal scale for the firm exists. For firms below the optimal scale the rate of growth of the firm increases as firm size decreases; the rate of growth of firms at the optimal scale is, of course, zero; the rate of growth of firms above the optimal scale is negative. A third possibility is that returns to scale are increasing. No optimal scale exists and the rate of growth of the firm increases with size. These qualitative results can easily be checked against a substantial body of evidence on rates of growth of individual firms and on the size distribution of firms.

On the basis of Bain's evidence on the character of economies of scale the most appropriate model of capital accumulation appears to be that based on constant returns to scale. For this model the rate

---

[62] See Bain [4], pp. 68–82.

[63] *Ibid.*, Table III, p. 72.

[64] *Ibid.*, pp. 61–62.

[65] *Ibid.*, pp. 83–93.

[66] [5], p. 99.

[67] See Lucas [45].

TABLE 6

AVERAGE FIRM SIZE, NUMBER OF FIRMS, AND AVERAGE FIRM GROWTH RATES FOR TWO-DIGIT INDUSTRIES IN THE U S

INDUSTRY GROUP	(1) NUMBER OF FIRMS	(2) AVERAGE FIRM SIZE BY QUARTILE ($MILLIONS)				(3) AVERAGE FIRM GROWTH RATE BY QUARTILE			
		FIRST	SECOND	THIRD	FOURTH	FIRST	SECOND	THIRD	FOURTH
FOOD	132	126.0	31.9	16.7	9 6	63	71	48	35
TEXTILE	83	69 4	24 0	12 4	8 5	82	50	59	37
PAPER	59	75 4	28 6	12.2	8 7	135	151	187	116
CHEMICALS	91	167 1	32.6	16 9	9 1	106	135	127	105
PETROLEUM	41	887.2	240.4	41 0	12 4	154	139	124	197
PRIMARY METALS	80	357 5	42 0	16 6	8 9	117	83	156	168
FABRICATED METALS	59	74 2	19.1	10.5	7 9	101	73	107	146
MACHINERY, EXCL ELECTRICAL	134	84 3	20 7	12 1	8 5	115	122	99	68
TRANSPORTATION EQUIPMENT	90	246 0	48.5	19 8	9 9	111	92	71	139

SOURCE HYMER AND PASHIGIAN [28], TABLES 1 AND 2, PP 558–59

of growth of the firm is independent of the size of the firm; a possible exception to this implication could occur for very small firms, where rates of growth higher than the average rate of growth might be expected. A careful study of growth rates of the 1,000 largest manufacturing firms (as of December 1946) has been made by Hymer and Pashigian.[68] Arithmetic average growth rates from 1946 to 1955 by two-digit industries are given in Table 6, together with the average size of firm in each quartile and the number of firms in the industry.[69] The findings are summarized as follows:

> Table [6] shows no relationship between the mean growth rate and size of firm. Mean growth rates are higher in the second quartile than in the first in four industries, and lower in six. Between the second and third quartiles the pattern is reversed: the mean growth rate rises in four industries and falls in six. In total the mean growth rate increases in fourteen cases and decreases in sixteen, suggesting a virtual equal probability of an increase or decrease.[70]

A similar but less complete study has been carried out by Meyer and Kuh, comparing large firms with very small firms.[71] The small firms are all smaller than the average size of firms in the fourth quartile of firms considered by Hymer and Pashigian. The large firms are larger than the average size of firms in the second quartile (in some industries the first quartile) of Hymer and Pashigian. Firms of intermediate size are omitted by Meyer and Kuh. Growth rates are measured over the period 1946–1950 rather than 1946–1955 as in the study of Hymer and Pashigian. The results show somewhat higher growth rates for small firms. These results are consistent with Bain's evidence on economies of scale and suggest the possibility of

---

[68] In [28].

[69] Average size of firm and number of firms are from Hymer and Pashigian [28], Table 1, p. 558; arithmetic averages of firm rates of growth are from Table 2, p. 559.

[70] *Ibid.*, p. 558.

[71] See [51], Chap. 10, esp. pp. 159–68.

TABLE 7

CONCENTRATION RATIOS FOR FOUR-DIGIT INDUSTRIES IN THE U S, 1947–1963

	1947	1954	1958	1963
FOUR LARGEST FIRMS	37.3	38.9	37.6	36.2
NUMBER OF INDUSTRIES	199	216	238	408
EIGHT LARGEST FIRMS	49.1	51.2	51.0	48.7
NUMBER OF INDUSTRIES	200	216	240	413
TWENTY LARGEST FIRMS	63.7	66.7	66.2	63.8
NUMBER OF INDUSTRIES	200	214	226	415

SOURCE. KAMERSCHEN [41], TABLE 6, P 241

increasing returns at small firm sizes. The evidence presented by Hymer and Pashigian suggests that these economies are limited to very small firms. Additional studies have been made by Mansfield and by Marcus.[72] Both are consistent with the findings of Hymer and Pashigian.[73]

Independence of the rate of growth of the firm and firm size implies that the relative size distribution of firms in a given industry should be stable over time. A number of alternative models of the size distribution of the firm have been proposed by Simon.[74] A complete theory of the size distribution of firms requires a theory of firm growth and a theory of entry and exit from an industry.[75] Since entry and exit occurs most frequently for firms of small size, a crude test of the stability of the size distribution can be made by analyzing changes in concentration ratios, changes in the proportion of industry shipments made by the largest four, eight, and 20 firms, over time. A summary of the evidence available for four-digit industries in manufacturing is given in Table 7;[76] concentration ratios averaged over four-digit manufacturing industries are very stable over the period 1947–1963. Over the whole period concentration ratios increased for 38 percent of the industry groups, remained the same for 22 percent, and decreased for 40 percent.[77] Our overall conclusion is that evidence on the rate of growth of firms by size and evidence on size distribution of firms is consistent with the hypothesis of constant returns to scale. Direct evidence on returns to scale from Bain's analysis of engineering estimates is consistent with constant returns to scale beyond a firm size equal to the minimum efficient size of plant.

---

[72] See [49] and [50], respectively.

[73] Mansfield [49], Table 2, p. 1032, reports regressions of the logarithm of size of firm at the end of period on the logarithm of size of firm at the beginning of period for periods approximately ten years in length for steel, petroleum, and tire industries. For firms of size greater than Bain's minimum efficient scale [4], growth is independent of firm size. Similar results are reported for nine industries for the period 1961–1964 by Marcus [50], Table V, p. 589.

[74] See Simon and Ijiri [60, 61, 62], and Simon and Bonini [59], see also Hymer and Pashigian [28]. Results for the alternative models vary in their implications for both mean and variance of the rate of growth. In the text we have considered only evidence on mean of the rate of growth cited by Hymer and Pashigian and by Meyer and Kuh [51].

[75] See Quandt [57].

[76] See Kamerschen [41], Table 6, p. 241. The concentration measures are ratios of the value of shipments for the four largest, eight largest, and 20 largest firms in each four-digit industry to the value of shipments for the industry.

[77] *Ibid.*, Table 4, p. 239.

TABLE 8

ESTIMATES OF THE ELASTICITY OF SUBSTITUTION, DISTRIBUTION PARAMETERS, AND THE SCALE PARAMETER FOR TWO-DIGIT MANUFACTURING INDUSTRIES IN THE U.S BY ZAREMBKA

INDUSTRY GROUP	(1) F-RATIO	(2) SCALE PARAMETER	(3) SECOND-ORDER PARAMETER
FOOD	1.93	1.44 (0.07)	0.23 (0.21)
TEXTILE	1.04	0.91 (0.04)	0.03 (0.13)
APPAREL	0.04	0.88 (0.06)	−0.06 (0.07)
FURNITURE	0.11	0.90 (0.06)	−0.44 (0.12)
PAPER	0.49	1.07 (0.03)	−0.02 (0.04)
CHEMICALS	6.30	1.01 (0.06)	0.00 (0.11)
PETROLEUM	0.82	1.11 (0.09)	0.02 (0.29)
STONE,CLAY, GLASS	2.02	0.96 (0.05)	0.00 (0.06)
PRIMARY METALS	1.18	1.05 (0.05)	0.12 (0.06)
FABRICATED METALS	1.65	0.98 (0.04)	−0.01 (0.06)
MACHINERY, EXCL ELECTRICAL	1.15	1.00 (0.07)	0.13 (0.08)
ELECTRICAL MACHINERY	0.18	1.02 (0.05)	0.07 (0.08)
TRANSPORTATION EQUIPMENT	1.35	1.12 (0.04)	−0.02 (0.04)

SOURCE ZAREMBKA [68], TABLE 2, P. 52

Our next objective is to consider evidence on economies of scale from econometric studies of the production function. This evidence may be divided into two groups: estimates based on direct estimation of the production function, and estimates based on the marginal conditions for labor and capital. The marginal conditions are appropriate for both constant and decreasing returns; for increasing returns these conditions are inappropriate since the corresponding second-order conditions are not satisfied. Evidence of increasing returns can be obtained only by direct estimation of the production function.

Estimates of the scale parameter of the CES production function by Zarembka, based on the second-order approximation in the logarithms of the variables given above,[78] are presented in Table 8. The statistical analysis proceeds in two stages. For each industry the hypothesis that the elasticity of substitution and the scale parameter are the same for the two years 1957 and 1958 is tested. Proceeding conditionally on the hypothesis that these parameters are the same, two hypotheses are tested: (1) that the scale parameter is equal to unity, and (2) that the elasticity of substitution is equal to unity. The hypothesis that the production functions are the same is rejected for only one industry group, Chemicals, at a level of significance of 0.025. In testing the remaining hypotheses, Scheffé's *S*-method[79] is

---

[78] See formula (5):

$$\ln Q = \ln \gamma + \nu\delta \ln K + \nu(1 - \delta) \ln L - \rho\nu\delta(1 - \delta)(\ln K - \ln L)^2.$$

[79] Under the hypothesis that the elasticity of substitution is equal to zero, the second-order parameter $\rho\nu\delta(1 - \delta)$ is equal to zero; see note 78 above. For discussion of the *S*-method, see Scheffé [58], pp. 66–72.

the appropriate statistical procedure. The two hypotheses are tested simultaneously at a level of significance of 0.025 by an appropriate choice of critical regions for each. The scale parameter is not significantly different from unity for any industry group; the elasticity of substitution is significantly different from unity for only one industry group, Furniture. The overall level of significance for the two sets of tests is 0.05 since the two sets of hypotheses are "nested."[80] The individual observations are establishment averages by state for each industry group so that Zarembka's results are consistent with the hypothesis of constant returns to scale at the plant level.

Estimates of the elasticity of substitution and the scale parameter based on the same second-order approximation in the logarithms of the variables have been made by Griliches for 1958 on the hypothesis that the production function is the same for all two-digit industries. Again employing Scheffé's *S*-method, the elasticity of substitution is not significantly different from unity, but the scale parameter is significantly larger than unity.[81] For a sample of 440 observations by state-region and two-digit industry for 1958 Griliches maintains the hypothesis that the elasticity of substitution is equal to unity for all industries. He fits a regression with distribution and efficiency parameters varying among industries and with the scale parameter the same for all industries. State-region dummy variables, representing variations in the efficiency parameter across state-regions, also enter the regression. For this regression the estimated value of the scale parameter is 1.039 with a standard error of 0.014; again, the scale parameter is significantly larger than unity.[82] The standard error of estimate is 0.0622 for this regression. He presents a regression with distribution parameters varying but efficiency and scale parameters the same across industries; this results in a standard error of estimate of 0.0782. The *F*-statistic for the hypothesis that the efficiency parameters are the same for all industries is 6.42; the null hypothesis can be rejected at almost any level of significance. Griliches also presents a regression with distribution, efficiency, and scale parameters the same for all industries; this regression contains no state-region dummy variables. The *F*-statistic for the null hypothesis that the efficiency and distribution parameters are the same for all industries and that state-region dummy variables are equal to zero is 6.96, which is also highly significant.[83]

We have reviewed three regressions reported by Griliches for a sample of 440 observations by state-region and two-digit industry. The evidence from these regressions clearly contradicts the hypotheses that efficiency and distribution parameters are the same for all industries and that efficiency parameters are the same for all state-regions. These findings are based on the maintained hypothesis that the elasticity of substitution is equal to unity for all industries and that the scale parameter is the same for all industries. They suggest that the hypothesis employed in Griliches' simultaneous estimates of the elasticity of substitution and scale parameters, that the production function is the same for all industries, is invalid. The hypothesis that distribution, efficiency, and scale parameters are the same for all industry groups is employed throughout Griliches' studies of economies of scale by size of plant. These studies purport to show that ". . . there may be no gains to be had from growing from a very small plant to a somewhat larger one. The economies of scale are largely in the medium- to large-scale plant range."[84] As Griliches notes, this finding does not survive the introduction of industry dummy variables, representing variations in efficiency across industries. The maintained hypothesis that distribution, efficiency, and scale parameters are the same for all industry groups is clearly in conflict with the empirical evidence, invalidating the conclusions based on the corresponding statistical model. Griliches' maintained hypothesis accounts for the conflict between his results and the evidence presented by Zarembka and between his findings on economies of scale at the plant level and those of Bain.

---

[80] *Ibid.*, esp. pp. 44–45 and 111–12.

[81] See Griliches [27], p. 297.

[82] *Ibid.*, Table 6, Regression 2.

[83] These regression results are presented by Griliches [27], Table 5, Regression 1, and Table 6, Regressions 1 and 2. The regression in Table 5 also includes an "age of capital" variable; however, the coefficient of this variable is almost equal to its standard error so that omitting "age of capital" from the regression would leave the standard error of the regression essentially unaltered.

Another approach to simultaneous estimation of scale and substitution parameters has been suggested by Dhrymes. Reverting to the original form (3) of the constant elasticity of substitution production function presented above, the marginal condition may be written

$$\frac{\partial Q}{\partial L} = \nu(1 - \delta)\gamma^{-\rho/\nu}Q^{(1+\rho/\nu)}L^{-(1+\rho)} = \frac{q_L}{q_Q}.$$

Since this relationship involves both the scale parameter ν and the substitution parameter ρ, the two parameters can be estimated simultaneously. Dhrymes fits this relationship to data on labor input and output for two-digit industries for 1958. For all but one industry group the scale parameter is slightly greater than unity, invalidating the marginal conditions employed by Dhrymes to estimate the scale parameter.[85] Equality between the marginal product of labor and the real wage is not a necessary condition for profit maximization with increasing returns. Dhrymes fails to observe that increasing returns invalidates his model.

Dhrymes' model has been employed by Eisner and Nadiri in estimating the scale and substitution parameters from investment functions. Their results are worth quoting in detail:

> From the general CES, two-factor function,
>
> $$Q = \gamma[\delta K^{-\rho} + (1 - \delta)L^{-\rho}]^{-\nu/\rho},$$

---

[84] Griliches [26], p. 155, argues: "One of the major objections to [this] . . . study is its implicit assumption that the coefficients of 'the' production function are the same for all two-digit industries. Such an objection is in part misdirected as far as the economies of scale and labor-quality measures are concerned since we are interested in estimating the *average* importance of these factors for *total* manufacturing." Griliches' estimates of the scale parameter with no allowance for variation in efficiency or distribution parameters across industries are appropriate averages only if industry dummy variables, associated with variations in efficiency parameters, and industry capital/labor ratios, associated with variations in distribution parameters, are uncorrelated with industry employment, associated with the scale parameter. Otherwise, estimates of the scale parameter are biased and do not represent an appropriate average.

[85] See [11], Table 2, p. 364.

> where ν = the degree of the function and $\rho = \frac{1}{\sigma} - 1$, we have
>
> $$\frac{\partial Q}{\partial K} = \delta\nu\gamma^{(1-\rho/\nu)}Q^{(1+\rho/\nu)}K^{-(1+\rho)}.$$
>
> Then, setting
>
> $$\frac{\partial Q}{\partial K} = \frac{c}{p}$$
>
> and writing $\delta\nu\gamma^{(1-\rho/\nu)}$ as $\beta^{1+\rho}$, we find that desired capital is
>
> $$K^* = \beta\left(\frac{p}{c}\right)^{\sigma}Q^{(\sigma+[1-\sigma]/\nu)}.$$
>
> For . . . $\sigma = 1$, this reduces to
>
> $$K^* = \alpha\frac{pQ}{c}.$$
>
> If $\nu = 1$, we of course have the constant returns case already noted. For $E_p = \sigma < 1$, and $\nu > 1$, however, we note that
>
> $$E_q = \sigma + \frac{1-\sigma}{\nu} = 1 + \frac{(1-\nu)(1-\sigma)}{\nu} < 1,$$
>
> and we have precisely the empirical results, which we observe. Indeed if one could really accept the marginal condition $\partial Q/\partial K = c/p$, one could use this investment function to estimate the degree as well as the elasticity of substitution of the CES production function. For then,
>
> $$\nu = (1 - E_p)/(E_q - E_p).$$
>
> Using the "best-fitting" estimates, $\hat{E}_p = .1306$ and $\hat{E}_q = .5873$ . . . we thus have $\nu = 1.84$. Taking the best-fitting estimates when the constant term is not deleted, $\hat{E}_p = .1576$, $\hat{E}_q = .8158$ and $\nu = 1.28$.[86]

Eisner and Nadiri, like Dhrymes, fail to observe that their interpretation invalidates their model. After this theoretical lapse was pointed out,[87] Eisner and Nadiri admit, "It is of course true that increasing returns are generally incompatible with competitive equilibrium."[88]

Bischoff has examined the original results of Eisner and Nadiri and has found that both the lag structure and the stochastic specification employed in their statistical model are defective. Eisner and Nadiri employ a distributed lag function with changes in the logarithms of capital as a dependent variable and lagged values of these changes as independent variables. They did not test for serial correlation of the residuals; least-squares estimates of the parameters of the distributed lag function corresponding to their models are inconsistent if the residuals are serially correlated.[89] The remaining variables in their model are changes in the logarithms of output and the ratio of the price of output to the price of capital services. They include seven lagged values of each whereas Bischoff includes eight lagged values. Bischoff also introduces an autoregressive parameter to generalize the stochastic structure of the Eisner-Nadiri model. These two changes produce a substantial improvement in the goodness of fit of the Eisner-Nadiri model. Estimating the generalized model by maximum likelihood, Bischoff reduces the sum of squared residuals from 0.5684×10^{-4} to 0.5043×10^{-4}.[90] Proceeding on the

---

[86] Eisner and Nadiri [16], p. 373, note 20.

[87] In Jorgenson and Stephenson [39], p. 348.

[88] [17], p. 217.

[89] See Griliches [23, 25].

[90] See [8], Table 1, p. 358, results labeled "c_1" are for the data employed by Eisner and Nadiri.

TABLE 9

ESTIMATES OF E_p, E_q, AND THE AUTOREGRESSIVE PARAMETER BY BISCHOFF

(1) $\hat{E}_p$	(2) $\hat{E}_q$	(3) AUTOREGRESSIVE PARAMETER	(4) SUM OF SQUARED RESIDUALS
0.1139	0.8945	1.0	1.7251×10^{-4}
1.0125	1.0277	0.0	1.0002×10^{-4}
1.0199	1.0294	−0.1	0.9984×10^{-4}
1.0203	1.0320	−0.2	1.0067×10^{-4}

SOURCE BISCHOFF [8], TABLE 3, P 364.

basis of the best-fitting specification of the lag structure, which generalizes that of Eisner and Nadiri, Bischoff has tested for the presence of serial correlation in the original Eisner-Nadiri lag specification. Bischoff finds that the residuals are serially correlated so that least-squares estimates of the parameters of the distributed lag function conditional on the Eisner-Nadiri stochastic specification are inconsistent.[91] Bischoff has corrected this error and has tested the hypothesis that the elasticity of substitution and the scale parameter in the Eisner-Nadiri model are equal to unity. These hypotheses cannot be rejected. The critical value of the appropriate F-statistic is 3.27; the calculated value is 2.08, well inside the interval for acceptance of the null hypothesis.[92] This analysis is based on quarterly data on investment in total manufacturing employed in the original Eisner-Nadiri article.[93]

Bischoff has estimated the elasticity of substitution and scale parameter for a different set of data, consisting of quarterly observations on U. S. postwar investment in producers' durables. An autogressive parameter is estimated together with the elasticities E_p and E_q defined by Eisner and Nadiri. The results are given in Table 9.[94] For the maximum likelihood estimate of the autoregressive parameter, the estimated value of E_p is 1.0199 and the estimated value of E_q is 1.0294. The values appropriate for a Cobb-Douglas function with constant returns to scale are both unity. The hypothesis that the autoregressive parameter is equal to zero is consistent with the evidence presented in Table 9. Employing this value of the autoregressive parameter as a maintained hypothesis, the estimated value of E_p is 1.0125 and the estimated value of E_q is 1.0277. Again testing the hypothesis that the elasticity of substitution and the scale parameter are equal to unity, Bischoff obtains an F-statistic of 0.419 with a critical value of 3.17.[95] The results provide no evidence against the Cobb-Douglas form of the production function with constant returns to scale. For completeness Bischoff has also tested the hypothesis that the elasticity of substitution is equal to zero for data on equipment investment. Allowing the scale parameter to be estimated, the resulting F-statistic is 26.91 with a critical value of 4.02; maintaining the hypothesis of constant returns to scale the F-statistic is 18.88 with a critical value of 3.17.[96]

---

[91] *Ibid.*, p. 359; this erroneous specification is employed by Eisner [14]; Eisner's estimates are also inconsistent.

[92] Bischoff [8], Table 2; results for Model No. 9 provide the appropriate test.

[93] See [16], p. 381. Eisner [14] has subsequently employed the original Eisner-Nadiri specification in analyzing this same body of data, ignoring Bischoff's demonstration that this specification is inferior to the one employed by Bischoff.

[94] See Bischoff [8], Table 3, p. 364; results in Table III-9 are with the constant term included in the regression.

[95] *Ibid.*, Table 4, p. 365; results for Model No. 9 provide the appropriate test.

Bischoff's findings effectively demolish the main conclusion of Eisner and Nadiri, namely, "Results . . . are generally consistent with the implications of CES production functions with elasticities of substitution nearer zero than unity and, possibly, increasing returns to scale."[97] A CES production function with increasing returns to scale invalidates the model used by Eisner and Nadiri, so that this conclusion is based on an internally contradictory interpretation of the evidence. The empirical results reported by Eisner and Nadiri have been traced by Bischoff to errors in specification of the lag structure and in the stochastic specification. When these errors are corrected, the empirical results are consistent with values of unity for the elasticity of substitution and the scale parameter.

Although Eisner and Nadiri appear to withdraw from their earlier interpretation of the empirical results and appear to acknowledge the validity of Bischoff's findings,[98] Eisner has subsequently employed the same interpretation in analyzing the data on equipment investment employed by Bischoff. Moreover, he has adopted the same stochastic specification used in the original work of Eisner and Nadiri, despite Bischoff's demonstration on the basis of evidence presented in Table 9 that this stochastic specification is inconsistent with the empirical facts. Eisner's interpretation of his results merits extensive quotation.

> First, working with a relation precisely analogous to that I used with Nadiri, and restricting ourselves to the quarterly data through 1962, we get a very direct confirmation of what was discovered in the [original] data. Our best estimate (in terms of minimizing the squares of residuals) indicates an elasticity of capital stock with respect to output of .68 and of capital stock with respect to relative prices . . . a nonsignificant −.08. When we take the data through 1966 our elasticity of capital stock with respect to output rises to .85 and that with respect to relative prices to .05.[99]

This "direct confirmation" as well as the original "discovery" are based on a specification of the lag structure and a stochastic specification Bischoff had already shown to be erroneous. Eisner continues:

> Perhaps I should have quit when I was ahead! But alas, Bischoff has himself led in another path and candor requires me to follow it as well, and to report my results, which in part concern his. Working with the independent variables defined as logarithms rather than differences in logarithms, our elasticity of capital stock with respect to output for the period through 1962 rises to .98 and that with respect to relative prices to a quite respectable .58! And when we operate through the 1966 period, the elasticity with respect to output goes to 1.03 and that with respect to relative prices to .78![100]

Eisner has published a series of five articles attempting to support the conclusion of Eisner and Nadiri that the elasticity of substitution is equal to zero and that returns to scale are increasing. After the errors in these papers had been exposed[101] and Eisner's own analysis of data on equipment investment, quoted above, had led to direct contradiction of the conclusions of Eisner and Nadiri, Eisner concludes: "Is [the] hypothesis [$E_p = E_q = 1$] confirmed after all, however unreasonable it may be to take it on faith without testing? My answer is to suggest that we remain agnostic."[102] Given the empirical evidence we have reviewed, the conclusions of Eisner and Nadiri must be rejected.

---

[96] *Ibid.*, Table 4, p. 365; results are for Model No. 5.

[97] [16], p. 381.

[98] Eisner and Nadiri [17], esp. Section II, p. 217, and Section V, pp. 219–20.

[99] [14], p. 59.

[100] *Ibid.*

Bischoff's examination of the findings of Eisner and Nadiri indicates that estimates of the elasticity of substitution from time series data depend critically on the correct specification of the lag structure of the underlying statistical model. By ignoring the serial correlation present in their model, Eisner and Nadiri produced a highly misleading set of results. When the errors in their specification are eliminated, their conclusions are reversed. A similar problem arises in reconciling estimates from time series and cross sections. In estimating the elasticity of substitution from time series careful attention must be paid to specification of the lag structure. Incorrect specifications lead to very substantial biases in the results. Ferguson has presented estimates of the elasticity of substitution from data on labor input and output for U. S. two-digit industries for the period 1949–1961 that are very similar to the cross section results.[103] Lucas and McKinnon have presented estimates for these industries that are substantially lower than the cross section estimates.[104] A careful investigation of the stochastic specification of the models employed in analyzing time series, analogous to Bischoff's investigation for estimates from time series data on capital input and output, remains to be done.

4. Optimal investment policy

■ Our objective in this paper has been to select an appropriate form for the production function and a characterization of returns to scale. We have examined the empirical evidence from studies of two-digit manufacturing industries in the United States for the postwar period. In examining the evidence we have taken the CES production function with a constant degree of returns to scale as a maintained hypothesis. This hypothesis is combined with an installation function for capital goods that generates a distributed lag investment function.

The CES production function may be described by four parameters: the elasticity of substitution, the degree of returns to scale, the distribution parameter, and the efficiency parameter. Except for the efficiency parameter, all of the parameters are assumed to be constant for cross section observations of a given industry by states or regions or for time series observations on a given industry. The efficiency parameter may vary over cross sections with effects removed by dummy variables for states or regions; alternatively, this parameter may vary over time with the effects removed by a time trend. We have reviewed evidence based on direct estimation of the production function and evidence based on marginal productivity conditions for labor and capital. Our overall conclusion is that the elasticity of substitution and the degree of returns to scale may be taken to be equal to unity; for these values of the parameters the CES production function reduces to the Cobb-Douglas production function with constant returns to scale.

---

[101] See Bischoff [8], Jorgenson and Hall [33], and Jorgenson and Stephenson [39].

[102] [14], p. 59.

[103] See [19].

[104] See [47] and [48], respectively.

Our description of technology incorporates costs of installation of investment goods. Under constant returns to scale the optimal policy for capital accumulation determines the rate of growth of capital and the labor/output and capital/output ratios corresponding to any set of prices of output and of labor and capital input. The marginal productivity condition for capital is incorporated into a distributed lag function determining investment expenditures. The desired level of capital is a perpetually moving target to which actual capital never converges. This characterization of optimal capital accumulation may be contrasted with the optimal policy under decreasing returns to scale. The distributed lag function determining investment depends on the long-run equilibrium demand for capital, which depends in turn on the prices of output and of labor and capital input. Under constant returns there exists no long-run equilibrium demand for capital as a function of prices.

The empirical evidence we have reviewed suggests that desired capital should be interpreted as a moving target rather than the long-run equilibrium value of capital. This interpretation is consistent with the neoclassical theory of investment behavior, employed extensively in econometric studies of investment expenditures.[105] In this theory the optimal policy for production and investment is described by an iterative process with output and employment determined by the marginal productivity condition for labor and the production function. Given output and employment, the desired level of capital is determined by the marginal productivity condition for capital; the actual level of capital is determined by a distributed lag investment function.[106] This policy is identical to that appropriate for a description of technology with production and installation subject to constant returns to scale. In econometric studies based on the neoclassical theory, the hypothesis that the elasticity of substitution is equal to unity has been employed.[107] This hypothesis is also consistent with the evidence we have reviewed.

The characterization of optimal investment policy has been the subject of considerable controversy in the econometric literature on investment. The problem is posed by Borch.

> (1) Capital stock is taken as fixed, and the firm determines the optimal level of output. This is a very reasonable short-run assumption. (2) The output level determined under the first assumption is taken as fixed, and the optimal amount of capital stock is determined. This seems to be a rather doubtful assumption. . . . It seems natural to assume that intelligent management will see more than one step ahead in this process and try to optimize output level and capital stock simultaneously.[108]

---

[105] In addition to the references given in notes 21 and 22 above, see Jorgenson and Stephenson [37, 38, 40], Jorgenson and Hall [31, 32, 33], Jorgenson and Siebert [35, 36], and Jorgenson and Handel [34].

[106] See Section 1 *infra.*

[107] See the references given in note 105 above.

[108] [9], p. 273.

The error in Borch's reasoning is that under constant returns to scale, the optimal program for capital accumulation involves selection of an optimal capital/output ratio followed by selection of an optimal rate of growth of capital. No matter how far ahead intelligent management can see, the optimal policy at each point in time may be described in this way. By suggesting that output and capital stock be optimized simultaneously, Borch assumes that optimal levels of output and capital exist so that returns to scale are decreasing. But this is contradicted by the empirical evidence we have reviewed.[109]

Borch's reasoning has been repeated and elaborated by Coen, Gould, Griliches, and Mundlak. Mundlak analyzes Koyck's flexible accelerator model, purporting ". . . to show how the use of current output as a measure of desired capital can lead to erroneous results."[110] Mundlak develops a model of the competitive firm as a basis for analyzing Koyck's distributed lag function. Long-run equilibrium for the model is characterized by optimal levels of output, labor, and capital input.[111] Mundlak assumes decreasing returns to scale; his analysis of Koyck's model presumes that long-run equilibrium is characterized by a supply function for output and demand functions for labor and capital input. Under constant returns to scale these supply and demand functions do not exist, so that Mundlak's analysis of Koyck's model is completely erroneous. This analysis is extended by Mundlak in developing a distributed lag investment model with desired capital a function of the ratios of input prices of output. While such a model would be appropriate under decreasing returns, the empirical evidence we have reviewed supports an assumption of constant returns. For this description of technology Mundlak's distributed lag investment model is inappropriate.

The arguments of Borch and Mundlak have been repeated by Griliches as part of a criticism of distributed lag investment models.

> In particular, the use of output as an exogenous variable in the demand equation needs further justification. There is no problem in treating it econometrically as predetermined, but it is hard to accept the long-run notions of "desired" capital being conditional on current output.[112]

Under constant returns the optimal policy is characterized by an optimal capital/output ratio and an optimal rate of growth of capital. The rate of growth of capital may be expressed as a function of the capital/output ratio. Alternatively, investment is determined by a distributed lag function that incorporates the marginal productivity of capital; under constant returns to scale the marginal product is a function of the capital/output ratio. A point of view similar to that of Griliches is expressed by Coen: "While output may be predetermined for the capital stock decision, it is not an exogenous variable for a profit maximizing firm. Like labor input and capital stock, output is a decision variable."[113] Coen is explicit in associating this point of view with decreasing returns to scale.[114] The error in Coen's analysis of distributed lag investment models is that constant returns to scale are excluded as a possible description of technology. From the evidence we have reviewed, constant returns provides the characterization of technology most appropriate to the analysis of investment behavior.

---

[109] See Section 3 *infra.*

[110] [54], p. 51, the reference is to Koyck [43].

[111] See Mundlak [53, 54].

[112] [22], p. 216.

[113] [10], p 370.

[114] *Ibid.*, p. 375.

The most elaborate presentation of the point of view of Borch, Coen, Griliches, and Mundlak on flexible accelerator models of investment has been given by Gould.[115] Gould's presentation is all the more remarkable in that it ostensibly relies on a theory of optimal capital accumulation similar to that we have outlined in Section 1. This theory is not used explicitly on the grounds that

> different adjustment mechanisms are derived when alternative assumptions are made about the form of the production function and the cost of adjustment function. We need not dwell here on the actual derivation of these alternative forms, since the main concern is not the adjustment mechanism *per se* but rather what happens when past or current sales are used in the specification of [desired capital] instead of the correct long-run equilibrium value.[116]

The error in this line of reasoning is more transparent in Gould's presentation than in those of Borch, Coen, Griliches, and Mundlak. Alternative assumptions about the form of the production function are critical to Gould's analysis. Under constant returns no "correct" long-run equilibrium value of capital exists. All of Gould's conclusions, like those of Borch, Coen, Griliches, and Mundlak, are predicted on the existence of such a long-run equilibrium value of capital. These conclusions are irrelevant to distributed lag models of investment based on constant returns to scale.

References

1. ALLEN, R. G. D. *Mathematical Analysis for Economists.* London: MacMillan & Co., 1938.
2. ARROW. K. J. "Optimal Capital Policy with Irreversible Investment," in J. N. Wolfe, ed., *Value, Capital, and Growth,* Chicago: Aldine Publishing Co., 1963, pp. 1–20.
3. ———, CHENERY, H. B., MINHAS, B., and SOLOW, R. M. "Capital-Labor Substitution and Economic Efficiency." *The Review of Economics and Statistics,* Vol. 43, No. 3 (August 1961), pp. 225–50.
4. BAIN, J. S. *Barriers to New Competition.* Cambridge, Mass.: Harvard Univ. Press, 1956.
5. ———. "Survival-Ability as a Test of Efficiency." *The American Economic Review,* Vol. 59, No. 2 (May 1969), pp. 99–104.
6. BELL, F. W. "A Note on the Empirical Estimation of the CES Production Function with the Use of Capital Data." *The Review of Economics and Statistics,* Vol. 47, No. 3 (August 1965), pp. 328–30.
7. ———. "The Role of Capital-Labor Substitution in the Adjustment of an Industry Across Regions." *Southern Economic Journal,* Vol. 31, No. 2 (October 1964), pp. 123–31.
8. BISCHOFF, C. W. "Hypothesis Testing and the Demand for Capital Goods." *The Review of Economics and Statistics,* Vol. 51, No. 3 (August 1969), pp. 354–68.
9. BORCH, K. "Topics in Economic Theory: Discussion." *The American Economic Review,* Vol. 53, No. 2 (May 1963), pp. 272–74.
10. COEN, R. N. "Tax Policy and Investment Behavior: Comment." *The American Economic Review,* Vol. 59, No. 3 (June 1969), pp. 370–79.
11. DHRYMES, P. J. "Some Extensions and Tests for the CES Class of Production Functions." *The Review of Economics and Statistics,* Vol. 47, No. 4 (November 1965), pp. 357–66.

---

[115] See [21].

[116] *Ibid.,* p. 589.

12. ———, and ZAREMBKA, P. "Elasticities of Substitution for Two-Digit Manufacturing Industries: A Correction." *The Review of Economics and Statistics*, Vol. 52, No. 1 (February 1970), pp. 115–17.
13. DOUGLAS, P. H. "Are There Laws of Production?" *The American Economic Review*, Vol. 38, No. 1 (March 1948), pp. 1–42.
14. EISNER, R. "Investment and the Frustrations of Econometricians." *The American Economic Review*, Vol. 59, No. 2 (May 1969), pp. 50–64.
15. ———. "Tax Policy and Investment Behavior: Further Comment." *The American Economic Review*, Vol. 60, No. 4 (September 1970), pp. 746–52.
16. ———, and NADIRI, M. I. "Investment Behavior and the Neo-Classical Theory." *The Review of Economics and Statistics*, Vol. 50, No. 3 (August 1968), pp. 369–82.
17. ———. "Neoclassical Theory of Investment Behavior: A Comment." *The Review of Economics and Statistics*, Vol. 52, No. 2 (May 1970), pp. 216–22.
18. EISNER, R. and STROTZ, R. "Determinants of Business Investment," in *Impacts of Monetary Policy*, Commission on Money and Credit, Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1963, pp. 59–337.
19. FERGUSON, C. E. "Time-Series Production Functions and Technological Progress in American Manufacturing Industry." *Journal of Political Economy*, Vol. 73, No. 2 (April 1965), pp. 135–47.
20. FUCHS, V. "Capital-Labor Substitution: A Note." *The Review of Economics and Statistics*, Vol. 45, No. 4 (November 1963), pp. 436–38.
21. GOULD, J. P. "The Use of Endogenous Variables in Dynamic Models of Investment." *Quarterly Journal of Economics*, Vol. 83, No. 333 (November 1969), pp. 580–99.
22. GRILICHES, Z. "The Brookings Model Volume: A Review Article." *The Review of Economics and Statistics*, Vol. 59, No. 2 (May 1968) pp. 215–34.
23. ———. "Distributed Lags: A Survey." *Econometrica*, Vol. 35, No. 1 (January 1967), pp. 16–49.
24. ———. "More on CES Productions." *The Review of Economics and Statistics*, Vol. 49, No. 4 (November 1967), pp. 608–10.
25. ———. "A Note on Serial Correlation Bias in Estimates of Distributed Lags." *Econometrica*, Vol. 29, No. 1 (January 1961), pp. 65–73.
26. ———. "Production Functions in Manufacturing: Some Additional Results." *Southern Economic Journal*, Vol. 35, No. 2 (October 1968), pp. 151–56.
27. ———. "Production Functions in Manufacturing: Some Preliminary Results," in M. Brown, ed., *The Theory and Empirical Analysis of Production*, Studies in Income and Wealth, Vol. 32, New York: Columbia Univ. Press, 1967, pp. 275–322.
28. HYMER, S. and PASHIGIAN, P. "Firm Size and Rate of Growth." *Journal of Political Economy*, Vol. 70, No. 6 (December 1962), pp. 556–69.
29. JORGENSON, D. W. "Anticipations and Investment Behavior," in J. Duesenberry, G. Fromm, L. Klein, and E. Kuh, eds., *The Brookings Quarterly Model of the United States*, Chicago: Rand McNally & Co., 1965, pp. 35–92.
30. ———. "Capital Theory and Investment Behavior." *The American Economic Review*, Vol. 53, No. 2 (May 1963), pp. 247–59.
31. ———, and HALL, R. E. "Application of the Theory of Optimum Capital Accumulation," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Amsterdam: North-Holland Publishing Co., 1971, pp. 9–60.
32. ———. "Tax Policy and Investment Behavior." *The American Economic Review*, Vol. 57, No. 3 (June 1967), pp. 391–414.
33. ———. "Tax Policy and Investment Behavior: Reply and Further Results." *The American Economic Review*, Vol. 59, No. 3 (June 1969), pp. 388–401.
34. JORGENSON, D. W. and HANDEL, S. S. "Investment Behavior in U. S. Regulated Industries." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 213–64.
35. JORGENSON, D. W. and SIEBERT, C. D. "A Comparison of Alternative Theories of Corporate Investment Behavior." *The American Economic Review*, Vol. 58, No. 4 (September 1968), pp. 681–712.
36. ———. "Optimal Capital Accumulation and Corporate Investment Behavior." *Journal of Political Economy*, Vol. 76, No. 6 (November/December 1968), pp. 1123–51.

37. Jorgenson, D. W. and Stephenson, J. "Anticipations and Investment Behavior in U. S. Manufacturing, 1947–60." *Journal of the American Statistical Association*, Vol. 64, No. 325 (March 1969), pp. 67–89.
38. ———. "Investment Behavior in U. S. Manufacturing, 1947–60." *Econometrica*, Vol. 35, No. 2 (April 1967), pp. 169–220.
39. ———. "Issues in the Development of the Neo-Classical Theory of Investment Behavior." *The Review of Economics and Statistics*, Vol. 51, No. 3 (August 1969), pp. 346–53.
40. ———. "The Time Structure of Investment Behavior in United States Manufacturing, 1947–60." *The Review of Economics and Statistics*, Vol. 49, No. 1 (February 1967), pp. 16–27.
41. Kamerschen, D. R. "Market Growth and Industry Concentration." *The Journal of the American Statistical Association*, Vol. 63, No. 321 (March 1968), pp. 228–41.
42. Kmenta, J. "On Estimation of the CES Production Function." *International Economic Review*, Vol. 8, No. 2 (June 1967), pp. 180–89.
43. Koyck, L. M. *Distributed Lags and Investment Analysis.* Amsterdam: North-Holland Publishing Co., 1954.
44. Leontief, W. W. *The Structure of the American Economy, 1919–1939.* 2nd ed. New York: Oxford Univ. Press, 1951.
45. Lucas, R. "Adjustment Costs and the Theory of Supply." *Journal of Political Economy*, Vol. 75, No. 4 (August 1967), pp. 321–35.
46. ———. "Optimal Investment Policy and Flexible Accelerator." *International Economic Review*, Vol. 8, No. 1 (February 1967), pp. 78–85.
47. ———. "Labor-Capital Substitution in U. S. Manufacturing," in A. C. Harberger and M. J. Bailey, eds., *The Taxation of Income from Capital*, Washington, D. C.: The Brookings Institution, 1969, pp. 223–74.
48. McKinnon, R. I. "Wages, Capital Costs, and Employment in Manufacturing: A Model Applied to 1947–58 U. S. Data." *Econometrica*, Vol. 40, No. 3 (July 1962), pp. 501–21.
49. Mansfield, E. "Entry, Gibrat's Law, Innovation, and the Growth of Firms." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1023–51.
50. Marcus, M. "A Note on the Determinants of the Growth of Firms and Gibrat's Law." *Canadian Journal of Economics*, Vol. 2, No. 4 (November 1969), pp. 580–89.
51. Meyer, J. and Kuh, E. *The Investment Decision.* Cambridge, Mass.: Harvard Univ. Press, 1957.
52. Minasian, J. R. "Elasticities of Substitution and Constant-Output Demand Curves for Labor." *Journal of Political Economy*, Vol. 69, No. 3 (June 1961), pp. 261–70.
53. Mundlak, Y. "Long-Run Coefficients and Distributed Lag Analysis." *Econometrica*, Vol. 35, No. 2 (April 1967), pp. 278–93.
54. ———. "On the Microeconomic Theory of Distributed Lags." *The Review of Economics and Statistics*, Vol. 48, No. 1 (February 1966), pp. 51–60.
55. Nerlove, M. *Estimation and Identification of Cobb-Douglas Production Functions.* Amsterdam: North-Holland Publishing Co., 1965.
56. ———. "Recent Empirical Studies of the CES and Related Production Functions," in M. Brown, ed., *The Theory and Empirical Analysis of Production*, Studies in Income and Wealth, Vol. 32, New York: Columbia Univ. Press, 1967, pp. 55–122.
57. Quandt, R. E. "On the Size Distribution of Firms." *The American Economic Review*, Vol. 56, No. 3 (June 1966), pp. 416–32.
58. Scheffé, H. *The Analysis of Variance.* New York: John Wiley and Sons, Inc., 1959.
59. Simon, H. A. and Bonini, C. P. "The Size Distribution of Business Firms." *The American Economic Review*, Vol. 48, No. 4 (September 1958), pp. 607–17.
60. Simon, H. A. and Ijiri, Y. "Business Firm Growth and Firm Size." *The American Economic Review*, Vol. 54, No. 1 (March 1964), pp. 77–89.
61. ———. "Effects of Mergers and Acquisitions on Business Firm Concentration." *Journal of Political Economy*, Vol. 79, No. 2 (March/April 1971), pp. 314–22.
62. ———. "A Model of Business Firm Growth." *Econometrica*, Vol. 35, No. 2 (April 1967), pp. 348–55.

63. Solow, R. M. "Capital, Labor, and Income in Manufacturing," in *The Behavior of Income Shares*, Studies in Income and Wealth, Vol. 27, Princeton, N. J.: Princeton Univ. Press, 1964, pp. 101–28.
64. Treadway, A. B. "On Rational Entrepreneurial Behaviour and the Demand for Investment." *Review of Economic Studies*, Vol. 36(2), No. 106 (April 1969), pp. 227–39.
65. Uzawa, H. "Time Preference and the Penrose Effect in a Two-Class Model of Economic Growth." *Journal of Political Economy*, Vol. 77, No. 4 (July/August 1969), pp. 628–52.
66. Walras, L. *Elements of Pure Economics*. Translated by W. Jaffé. Homewood, Ill.: Richard D. Irwin, Inc., 1954.
67. Wicksell, K. "A Mathematical Analysis of Dr. Åkerman's Problem," in K. Wicksell, *Lectures on Political Economy*, edited by L. Robbins, translated by E. Classen, Vol. I, London: Routledge and Kegan Paul, 1934, pp. 274–99.
68. Zarembka, P. "On the Empirical Relevance of the CES Production Function." *The Review of Economics and Statistics*, Vol. 52, No. 1 (February 1970), pp. 47–53.
69. ———, and Chernicoff, H. "Further Results on the Empirical Relevance of the CES Production Function." *The Review of Economics and Statistics*, Vol. 53, No. 1 (February 1971), pp. 106–10.

Pricing pollution and other negative externalities

Richard A. Tybout

Professor of Economics
The Ohio State University

Equilibrium conditions are described for pollution or "externalities of production" treated as substitute products. The results are contrasted with those based on traditionally assumed fixed proportions between product and waste outputs. The "neutrality of bribery or compensation" argument is refuted. With linear homogeneous production functions, compensation leads to exhaustion of product; bribery does not. Only certain selected production functions and conditions can lead to positive aggregate profits in a bribe-paying industry. Requirements are described. The analysis is applied to a number of related issues: the blackmail problem, third party pricing, and public goods aspects.

1. Introduction

■ Discommodities are generally externalities. The incentive is to disown rather than to own, to avoid rather than to ration. But avoiding and disowning are becoming more difficult as population and GNP grow. So it is that economic theory faces a problem of increasing magnitude: the allocation of negative externalities.

Intuitively, one would expect all of the familiar rules of market operation to apply to discommodities, except in reverse. Supply should be negatively sloped. Demand should be positively sloped. And the originator of a discommodity should pay the receiver to take it off his hands. But economic theory is not so constructed at the moment. An imposing part of received theory on the economics of pollution asserts that legal responsibility can be assigned to either the pollution originators or to those damaged without affecting the allocation of resources. Either way, it is held, the parties could negotiate a monetary settlement that would lead to identical waste output in the absence of transactions costs.

The argument originated with R. H. Coase and has been developed in a series of supporting and qualifying articles.[1] It has recently appeared in what is perhaps the leading scholarly monograph

---

Richard A. Tybout received the B.Ch.E. degree from the University of Delaware (1943), the M.S.E. in chemical engineering (1947) and the M.A. in economics (1949), both from the University of Michigan, and the Ph.D. in economics (1952), also from that institution. Professor Tybout served as a price economist with the Office of Price Stabilization from 1951–1952 and has held professorial positions at The Ohio State University since 1954. He has published extensively on subjects ranging from transportation economics to environmental problems.

The author wishes to thank the U. S. Office of Water Resources Research for financial assistance and J. Hayden Boyd, Allen V. Kneese, and Robert E. Kuenne for review of an earlier version of the manuscript. Sole responsibility for the results rests, of course, with the author.

[1] See Coase [5]; and Buchanan and Stubblebine [4], Davis and Whinston [7], and Wellisz [15].

on the economics of pollution.[2] As noted above, neutrality is claimed only in the absence of transactions costs. This disclaimer limits consideration to private goods, and the argument will be evaluated in that context, though a final part of the present article will extend the theory developed here into public goods. A second limitation, as often implicit as explicit, is that indirect effects of the distribution of income on demand are ignored.[3] In this sense, the analysis is conducted in a partial equilibrium context. Finally, we shall be concerned with only unilateral, as opposed to multilateral, externalities.[4]

Others[5] have pointed to certain anomalies in the Coase argument, though their criticisms imply an alternative kind of symmetry, i.e., that compensation and bribery are equally workable market alternatives. The objection they make is to that part of the thesis that predicts the same allocation of resources with either payment alternative.

The main conclusion of the present paper is that the transaction for internalizing a negative externality cannot, in general, run either way. We shall attempt to show that viable market operation requires that an internalizing payment must in general take the form of compensation running from the discommodity originator to the damaged party. A transaction in the opposite direction, known as bribery, cannot exist except in rare cases dependent on fortuitous circumstances. Section 2 presents an equilibrium formulation with the externality treated as a substitute product. Section 3 summarizes and restates in a more general way the argument for neutrality of compensation and bribery, or the "Coase Theorem," as we shall call it. Section 4 gives the analytical basis for the author's conclusion that bribery and compensation are far from symmetrical. Section 5 deals with a number of related issues, largely theoretical and largely raised by discussions in the literature on the Coase Theorem.

The theoretical analysis and the conclusions of the present work appear in the context of pollution control. Nevertheless, it is hoped that the analysis of bribery may be suggestive for such diverse topics as the economics of crime and the logic of the agricultural soil bank program. The reference is to the special circumstances that surround a situation where payment is made for not doing something, as described in the present analysis of bribery.

2. Economics of internalization

■ A standard treatment of multiproduct commodity pricing provides the basis for the theory of internalization. For positive externalities, the results are the same as in conventional commodities markets, though the externalities literature has not been so developed. For negative externalities, the results are unfamiliar.

A general-purpose statement of the optimization problem can be made with externality J, either a commodity or a discommodity, to be internalized. The originator, A, of J sees the classical production problem as

---

[2] See Kneese and Bower [11], pp. 98–109.

[3] The effects of compensation and bribery are, of course, not neutral when the distribution of income affects output decisions. For a discussion of this case, see Dolbear [8].

[4] For discussion of some of the problems involved with multilateral externalities, see Davis and Whinston [6].

[5] See Boyd [2], Bramhall and Mills [3], and Wellisz [15].

$$\text{Maximize} \quad \pi_A = P_A X_A \pm P_J X_J - wL_A - rC_A \tag{1}$$

$$\text{Subject to} \quad X_A - F(X_J,L_A,C_A) = 0\,. \tag{2}$$

The receiver, *B*, of *J* sees it as

$$\text{Maximize} \quad \pi_B = P_B X_B \mp P_J X_J - wL_B - rC_B \tag{3}$$

$$\text{Subject to} \quad X_B - G(X_J,L_B,C_B) = 0\,. \tag{4}$$

Conventional products *A* and *B* are produced in quantities X_A and X_B and sold for prices P_A and P_B. There are two inputs, labor, *L*, and capital, *C*, priced at wage rate *w* and interest rate *r*. A quantity X_J of the externality *J* is produced by *A* and received by *B*. The internalization price is P_J. The sign convention before the *J* variable is to be interpreted symmetrically. Thus, if *J* is a positive externality, the plus sign in (1) and the minus sign in (3) apply. If *J* is a negative externality, the signs are reversed. We assume that prices are viewed parametrically by all parties and constant returns to scale obtain throughout. Second-order conditions are those required for quasi-concavity.

Internalization takes place through the establishment of markets that produce a positive price P_J in place of the pre-internalization price $P_J = 0$. J. E. Meade's celebrated example of externality, the production of honey from nectar in apple blossoms, provides an example of a positive externality.[6] Interpret commodity *A* as apples, commodity *B* as honey, and externality *J* as nectar; $J \equiv N$, with the plus sign before the second term on the right hand side of (1) and a minus in the same position in (3). Internalization then produces the familiar results

$$P_N = -P_A \frac{\partial X_A}{\partial X_N} \tag{6}$$

$$P_N = P_B \frac{\partial X_B}{\partial X_N}\,. \tag{7}$$

One conspicuous difference is that Meade designed his analysis to deal with indirect as well as direct effects. The latter are the only effects of interest here. More important from the standpoint of externalities, however, is the way in which the commodity nectar is treated. Meade introduced no such commodity explicitly. Using our notation, he employed the following production functions in place of our (2) and (4):

$$X_A - H_A(L_A,C_A) = 0 \tag{8}$$

$$X_B - H_B(X_A,L_B,C_B) = 0\,. \tag{9}$$

In effect, (8) and (9) require that apples and nectar be produced as strictly joint products, i.e., $X_N = kX_A$ and $\frac{\partial X_N}{\partial X_A} = k$.[7] The present

---

[6] See [13].

[7] Meade finds a tax needed on honey production in the amount (using our notation) of:

$$\text{Tax} = \frac{\partial X_B}{\partial X_A}\,\frac{X_A}{X_B}X_B P_B = \frac{\partial X_B}{\partial X_A}X_A P_B.$$

In our formulation, the tax is

formulation, of course, allows the conventional commodities and the externality to be viewed as substitute products.

Now, reverse the externality and consider the case of pollution. Interpret J as P and use the minus sign before P_P in (1) and the plus sign before P_P in (3). These sign changes now state that the polluter, industry A, must pay a price equal to P_P for each unit of pollution produced. The price (effluent tax) is paid to industry B for damages. We can imagine that industry A is located upstream of industry B. The former produces commodity A and pollution P. The latter produces commodity B and involuntarily receives pollution in the course of using river water. Pollution has a negative effect on the production of commodity B. The first-order conditions of maximization are the same as before except for the following new internalization price conditions:

$$P_P = P_A \frac{\partial X_A}{\partial X_P} \tag{10}$$

$$P_P = -\,P_B \frac{\partial X_B}{\partial X_P}\,. \tag{11}$$

Compare equations (6) and (7). Since prices are everywhere positive, the marginal production conditions must be of opposite sign in (10) and (11) from what they were in (6) and (7). Our intuition is further confirmed by analyzing supply and demand conditions in the neighborhood of equilibrium. See the summary of production and market conditions in Tables 1 and 2.[8]

$$P_N X_N = X_N P_B \frac{\partial X_B}{\partial X_N}.$$

The two formulations are equivalent if

$$\frac{X_N}{X_A} = \frac{\partial X_N}{\partial X_A}$$

and this will be the case when

$$\frac{X_N}{X_A} = \frac{\partial X_N}{\partial X_A} = k.$$

[8] The first-order conditions for maximization of equation (1) subject to (2) derived using a Lagrangean multiplier λ are:

$$-w + \lambda \frac{\partial F}{\partial L_A} = 0 \tag{i}$$

$$-r + \lambda \frac{\partial F}{\partial C_A} = 0 \tag{ii}$$

$$\pm P_J + \lambda \frac{\partial F}{\partial X_J} = 0 \tag{iii}$$

$$P_A - \lambda = 0 \tag{iv}$$

$$X_A - F(X_J, L_A, C_A) = 0. \tag{v}$$

The effect of a small change in P_N is found by solution of the following system (where superscripts designate partial differentiation), based on equations (i)

These results were, of course, all implicit in the statement of the internalization problem in equations (1) through (4), but this statement was not arbitrary. The results conform to the underlying utility relationships one would use to define commodities (positive externalities) and discommodities (negative externalities). Thus, in output A–output J space, there are indifference curves with slope

$$\left(\frac{\partial X_A}{\partial X_J}\right)_{U_0} = -\frac{(\partial U/\partial X_J)}{(\partial U/\partial X_A)}. \tag{12}$$

Since A is a conventional commodity, $\dfrac{\partial U}{\partial X_A} > 0$. When $J \equiv N$, $\dfrac{\partial U}{\partial X_J} > 0$. When $J \equiv P$, $\dfrac{\partial U}{\partial X_J} < 0$. Compare the production conditions for industry A in Table 1. Analogous relationships can be defined in output B–input J space.[9]

Internalization of the negative externality was conducted above by the process of compensation. Bribery internalizes by making the

---

through (v):

$$\begin{bmatrix} F^{L_AL_A} & F^{L_AC_A} & F^{L_AX_N} & 0 & F^{L_A} \\ F^{C_AL_A} & F^{C_AC_A} & F^{C_AX_N} & 0 & F^{C_A} \\ F^{X_NL_A} & F^{X_NC_A} & F^{X_NX_N} & 0 & F^{X_N} \\ 0 & 0 & 0 & 0 & -1 \\ F^{L_A} & F^{C_A} & F^{X_N} & -1 & 0 \end{bmatrix} \begin{bmatrix} \dfrac{\partial L_A}{\partial P_N} \\ \dfrac{\partial C_A}{\partial P_N} \\ \dfrac{\partial X_N}{\partial P_N} \\ \dfrac{\partial X_A}{\partial P_N} \\ \dfrac{\partial \lambda}{\partial P_N} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \\ 0 \end{bmatrix}.$$

Let $|D|$ denote the determinant formed from the matrix of coefficients. Then, $\dfrac{\partial X_N}{\partial P_N} = \dfrac{-|D_{33}|}{|D|}$. Quasi-concavity implies that the signs of unbordered Hessians follow the rule $(-1)^s$, where s is the number of columns or rows. See Kuenne [12], pp. 181–82. The above matrix expanded by cofactors gives the third-order principal minor preceded by a negative sign: $|D| = -|D_{44,55}|$, which is an unbordered Hessian following the same sign convention. Therefore $|D| > 0$. The same sort of expansion gives $|D_{33}| = -1|D_{33,44,55}| < 0$. Hence $\dfrac{\partial X_N}{\partial P_N} > 0$. This is, of course, the slope of a supply curve for a conventional commodity in the neighborhood of equilibrium. By the use of symmetry, other results can be quickly derived from the same mathematical framework. Thus, with $J \equiv P$, the negative sign applies in (iii) and the (-1) is replaced by a $(+1)$ in the solutions vector, thus giving the slope of the pollution supply curve $\dfrac{\partial X_P}{\partial P_P} < 0$. Industry B can be handled in the same way. Replace F with G and A with B, then proceed as before. The sign convention in (iii) is reversed to give slopes of the demand curves $\dfrac{\partial X_N}{\partial P_N} < 0$ and $\dfrac{\partial X_P}{\partial P_P} > 0$ in industry B.

[9] With X_J as an input, it is necessary to solve (12) for X_J's marginal utility:

$$\frac{\partial U}{\partial X_J} = -\frac{\partial U}{\partial X_A}\left(\frac{\partial X_A}{\partial X_J}\right)_{U_0}. \tag{i}$$

Substitute (i) in a relationship analogous to (12) to obtain:

$$\left(\frac{\partial X_B}{\partial X_J}\right)_{U_0} = -\frac{\partial U/\partial X_J}{\partial U/\partial X_B} = \frac{\partial U/\partial X_A}{\partial U/\partial X_B}\left(\frac{\partial X_A}{\partial X_J}\right)_{U_0}. \tag{ii}$$

TABLE 1

OPTIMIZATION CONDITIONS

COMMODITY	INDUSTRY			
	A		B	
NECTAR	$\frac{\partial X_A}{\partial X_N} < 0$	SUBSTITUTE PRODUCTS P_N PRECEDED BY +	$\frac{\partial X_B}{\partial X_N} > 0$	INPUT AND OUTPUT INCREASE TOGETHER, P_N PRECEDED BY –
POLLUTION	$\frac{\partial X_A}{\partial X_P} > 0$	JOINT PRODUCTS P_P PRECEDED BY –	$\frac{\partial X_B}{\partial X_P} < 0$	INPUT AND OUTPUT ARE SUBSTITUTES P_P PRECEDED BY +

negative of the negative externality a commodity that is internalized in the same way as a positive externality. The problems which this creates will be explored by reference to the mechanism envisaged by Coase and others.

3. The Coase Theorem

■ The following presentation draws on a graphical exposition by Turvey, later adopted by Kneese and Bower.[10] The last two authors qualify their acceptance of the theorem in a long-run context, as we shall note, but not for the short run.

As in Section 2, above, two industries, *A* and *B*, are assumed to be producing conventional outputs *A* and *B* as well as (in the case of *A*) producing pollution and (in the case of *B*) being constrained by *A*'s production of pollution. Coase discusses the process in the context of bilateral monopoly, whereas our parametric treatment of prices is most consistent with universal pure competition. The situation with bilateral monopoly will be noted at a later point.

TABLE 2

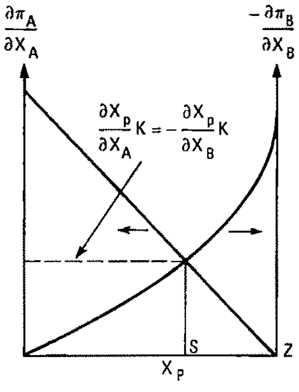
INTERNALIZATION MARKETS

COMMODITY	INDUSTRY	
	A (SUPPLY)	B (DEMAND)
NECTAR	$\frac{\partial X_N}{\partial P_N} > 0$	$\frac{\partial X_N}{\partial P_N} < 0$
POLLUTION	$\frac{\partial X_P}{\partial P_P} < 0$	$\frac{\partial X_P}{\partial P_P} > 0$

Consider Figure 1. The negatively sloped line is a supply curve for pollution expressed indirectly through marginal profits on output *A*. Outputs *A* and *P* are simultaneously expanding. If the expansion takes place with strict jointness in the production of *A* and *P*, there is a unique downward-sloping supply curve of pollution. If the output takes place with variable proportions of *A* and *P*, then the pollution supply curve depends on both prices P_A and P_B. See Tables 1 and 2.

FIGURE 1

COMPENSATION EQUILIBRIUM



---

[10] See Turvey [14] and Kneese and Bower [11], p. 100.

We shall allow X_A and X_P to vary independently but assume that the single downward-sloping curve in Figure 1 represents prices consistent with the processes to be described—in particular, with a compensation equilibrium at point S. $P_P = 0$ at point Z by construction. Thus, industry A is in equilibrium at Z without internalization.

The upward-sloping line indicates increasing marginal losses to B as output P is expanded and, by hypothesis, received by B. This is a demand curve for pollution with a compensation regime. See Tables 1 and 2. Variable proportions of input P and output B are possible; therefore, it is necessary to assume prices P_P and P_B such that the compensation equilibrium is reached at point S and there is no internalization of pollution at point Z. Industry B is assumed, for the time being, to have no transferable resources.

The units on the two vertical axes are adjusted so that equal vertical distances give equal dollar quantities. The production conditions of Section 2 are carried over so that both industries A and B have linear homogeneous production functions. At the compensation equilibrium, $X_P = S$ and we shall let $P_P = K$. Then by virtue of linear homogeneity, there is complete exhaustion of product:

$$\pi_A = P_A X_A - KS - wL_A - rC_A = 0 \tag{1a}$$

$$\pi_B = P_B X_B + KS - wL_B - rC_B = 0 . \tag{3a}$$

The bribery mechanism can be best understood with a supply and demand interpretation. To the right of $X_P = S$, industry B will minimize losses by paying A a bribe not to pollute. The bribe will give A more income than would the production of commodity A for sale in the market (with simultaneous production of P). To the left of $X_P = S$, industry B's losses are not as great as the size of the bribe that would be required to induce A to withhold production. Hence, equilibrium results again at $X_P = S$. Note that B's curve is now a demand for abatement, downward sloping from right to left, and A's curve is a supply of abatement, upward sloping from right to left.

The essential assumption in the argument is that pollution abatement can be treated as a commodity to be internalized. Call this commodity "withheld pollution," or "withholding," W. Then $J \equiv W$ and all the same marginal conditions apply as with $J \equiv N$. Assume for the time being that the level of output from which pollution will be withheld is known and accepted by all parties. This is the noninternalized level $X_P = Z$, which obtains when $P_P = 0$. Then

$$X_W = Z - X_P . \tag{13}$$

Since Z is a given constant,

$$\frac{dX_W}{dX_P} = -1 . \tag{14}$$

Internalization by bribery can be produced by commodity W in the general model with $J \equiv W$, a plus sign before P_W in (1), and a negative sign before P_W in (3) to give

$$P_W = -P_A \frac{\partial X_A}{\partial X_W} = P_B \frac{\partial X_B}{\partial X_W} . \tag{15}$$

Now, as a consequence of (14), our previous equilibrium with P_P can be turned into an equilibrium with P_W:

$$P_W = P_A \frac{\partial X_A}{\partial X_P}\bigg|_{X_P=S} = -P_B \frac{\partial X_B}{\partial X_P}\bigg|_{X_P=S} = P_P = K. \qquad (16)$$

The marginal conditions of general equilibrium with "withholding" are identical but of opposite sign to those with pollution at the same level of output $X_P = S$. Total profits with bribery are

$$\pi_A = P_A X_A + K(Z - S) - wL_A - rC_A = KZ \qquad (1b)$$

$$\pi_B = P_B X_B - K(Z - S) - wL_B - rC_B = -KZ. \qquad (3b)$$

Writers in the Coase tradition typically hold that the lump sum transfer KZ shown in (1b) and (3b) as compared with (1a) and (3a) is of distributional significance only (ignoring, as we do, indirect effects on the structure of demand).

4. Critical evaluation

■ A statement of the total-profits anomaly appears in Bramhall and Mills.[11] Referring to industry *B*, they hold:

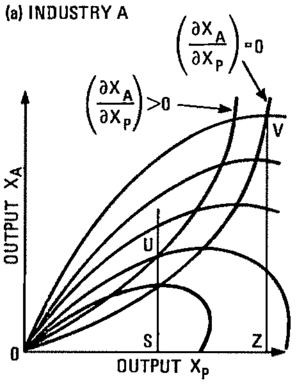
> Under the payments scheme [compensation], profits will be larger than they would have been in the absence of intervention, and under the fees scheme [bribery] profits will be smaller than in the absence of intervention. On the usual assumptions about entry and exit, entry will take place in the former case and exit in the latter case.

Wellisz[12] had previously made the point that for the Coase mechanism to hold, it is necessary to postulate nontransferable resources with Ricardian rents. We shall investigate the implications of transferable resources.

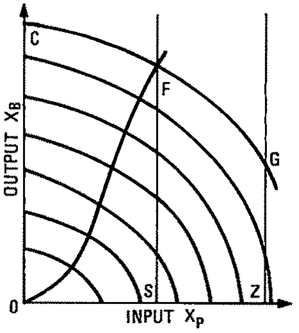
FIGURE 2

PRODUCTION–EXPANSION CURVES

(a) INDUSTRY A



(b) INDUSTRY B



☐ **The compensation case.** Production-expansion curves are shown in Figure 2 for the compensation case. The purpose of the curves is to show that the adjustment process is smooth, continuous, and follows conventional economic principles with compensation, in contrast to the situation we shall find with bribery. Figure 2 also gives information useful for the analysis of the bribery case and, in fact, can be reinterpreted for a bribery regime, though most of such reinterpretation will be left for the reader after the difficulties of bribery have been presented.

Figure 2(a) gives transformation curves (concave to the origin) for outputs *A* and *P* in industry *A*, each curve subject to a fixed resource constraint, which, in this case, is a real fund available for capital and labor at wage rates and interest rates fixed elsewhere in the economy. The transformation curves include positively sloped segments which imply that commodity *A* and pollution *P* outputs expand simultaneously over a part of the range. If there were not a positively sloped segment of each curve, there would be no pollution output in the first place. With strictly joint products, the transformation curves would collapse to a single ray from the origin, each point along which would correspond to different resource inputs, as well as different outputs. The slopes of the transformation curves at each point correspond to price ratios and represent isoprofit loci.[13]

---

[11] In [3], p. 616.

[12] In [15].

Two expansion curves are shown, one labeled $\left(\frac{\partial X_A}{\partial X_P}\right) = 0$, the other labeled $\left(\frac{\partial X_A}{\partial X_P}\right) > 0$. Since industry A obtains revenue from output A, but not from P, the only points of interest are on the positively sloped parts of the transformation curves to the left of the expansion curve $\left(\frac{\partial X_A}{\partial X_P}\right) = 0$. This last represents, of course, the expansion path that industry A would take in the absence of any charge assessed for pollution: $P_P = 0$. Compare equation (10). To relate Figure 2(a) to Figure 1, the equilibrium points $X_P = Z$ and $X_P = S$ are shown. These correspond to points V and U, respectively, in Figure 2(a).

Figure 2(b) is constructed in the same way. The curves concave to the origin show transformation trade-offs of output B and input P with alternative quantities of fixed total resources. There are no positively sloped parts of these curves. Input P retards the production of output B and the latter retards the assimilation of P, each over the entire range. A possible expansion path of industry B to point F, the compensation equilibrium, is shown. With demand and supply fixed, the transformation curves are also isoprofit curves with slopes equal to price ratios.[14] Looking back to the origin from point F, the

---

[13] The slopes of these curves can be found as follows. First, note that the marginal profits from commodity A are given by

$$\frac{\partial \pi_A}{\partial X_A} = P_A - P_P\frac{\partial X_P}{\partial X_A} - w\frac{\partial L_A}{\partial X_A} - r\frac{\partial C_A}{\partial X_A}. \qquad \text{(i)}$$

If we define

$$\alpha_A = P_A - w\frac{\partial L_A}{\partial X_A} - r\frac{\partial C_A}{\partial X_A}, \qquad \text{(ii)}$$

it can be shown that

$$\lim_{\alpha_A \to P_P \frac{\partial X_P}{\partial X_A}} d\pi = 0. \qquad \text{(iii)}$$

Rearranging (i) with zero marginal profits then gives

$$\frac{\partial X_P}{\partial X_A} = \frac{P_A - w\dfrac{\partial L_A}{\partial X_A} - r\dfrac{\partial C_A}{\partial X_A}}{P_P}, \qquad \text{(iv)}$$

which expresses the slopes of the production transformation curves at all equilibrium values of the indicated parameters and marginal conditions. Since $wL_A + rC_A$ is constant for any given contour of Figure 2(a), $\frac{\partial X_P}{\partial X_A} = \frac{dX_P}{dX_A}$ and the slopes shown in (iv) are reciprocals of the slopes shown in Figure 2(a).

[14] The slopes of the transformation curves shown in Figure 2(b) can be understood by reference to exactly the same kind of relationships described in note 13, above. Thus, marginal profits from commodity B are given by

$$\frac{\partial \pi_B}{\partial X_B} = P_B + P_P\frac{\partial X_P}{\partial X_B} - w\frac{\partial L_B}{\partial X_B} - r\frac{\partial C_B}{\partial X_B}. \qquad \text{(i)}$$

and by exactly analogous reasoning, it can be shown that with zero marginal

$$\frac{\partial X_P}{\partial X_B} = \frac{-P_B - w\dfrac{\partial L_B}{\partial X_B} - r\dfrac{\partial C_B}{\partial X_B}}{P_P}. \qquad \text{(ii)}$$

transformation contours represent higher and higher levels of marginal and average profits that could be achieved by a contraction of the industry with parametric increases in prices P_B and P_P.

No problems of adjustment are presented by the expansion curves in a compensation regime. Industry A can simultaneously expand or contract X_A and X_P in all combinations. Industry B can simultaneously expand the output of X_B and the input of X_P, for both of which it receives compensation along a common expansion curve. A reduction in demand consists of moving the positively sloped curve to the left in Figure 1; an expansion, to the right. Symmetrical movements of the negatively sloped supply curve produce expected results in the compensation price. Stability conditions are conventional, as noted in Section 2.

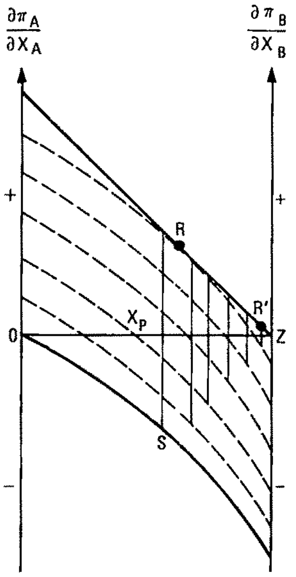
☐ **The bribery case.** Consider the situation with bribery. Industry A may expand, but it is not clear what will be produced. With Z held constant, $\frac{\partial X_A}{\partial X_W} < 0$ is a property of all expansion curves for industry A. See Figure 2(a), which is consistent with $\frac{\partial X_A}{\partial X_P} > 0$ and equation (14). One solution is to drop equation (14) and assume that witholding is sold from some level of $X_P > Z$. Whether this is a possible result depends on B's willingness to believe that A would, in fact, produce pollution at a level greater than Z. Point Z represents A's noninternalization optimum. Expansion in industry A might take place by balancing losses from A and W at the margin until the total profits in industry A are zero, but economic theory gives us no guidance as to whether or by what expansion path this will be done. In Figure 2(a), marginal profits are zero at point U, though total profits are KZ. See equations (15) and (1b).

The difficulty arises from a conflict between marginal and total profits. Marginal profits are zero at $X_P = S$ with bribery because industry A is being paid the marginal cost of curtailing pollution which, even with variable proportions, means a reduction of output A. Coase emphasized marginal profits; Bramhall-Mills, total profits. Total profits cannot help influencing the adjustments, but when there is a conflict between total and marginal profits, the outcome is a behavioral question. Stability is not assured, even if we assume that B has infinite resources with which to pay bribes.

Now, let us assume away instability in industry A. Assume that new competition is prevented by one means or another from entering industry A and that demand and cost conditions remain as previously given for A. Long-run contraction of X_B will raise the price P_B and result in higher marginal profits $\frac{\partial \pi_B}{\partial X_B}$. A bribery equilibrium might be established if P_B is raised high enough to offset the lump sum transfer. Whether this will happen depends on a complex set of circumstances. See Figure 3, which is a reproduction of Figure 1 with the right-hand axis reversed in sign. The left-hand axis remains as

FIGURE 3

BRIBERY EQUILIBRIUM



With $wL_A + rC_A$ constant, $\frac{\partial X_P}{\partial X_B} = \frac{dX_P}{dX_B}$, and hence the slopes in (ii) are the reciprocals of those shown in Figure 2(b).

before. The dotted lines show upward shifts in marginal profits to industry B as contraction takes place.

A total profits equilibrium is reached by B if some positive level of marginal profits results in zero total profits with bribery. Such an equilibrium is shown at point R. The reader can verify that if the curves were folded over with negative profits increasing upward as in Figure 1, the supply-demand "equilibrium" would move down the curve $\frac{\partial \pi_A}{\partial X_P}$ versus X_P. The vertical lines indicate points of bribery "equilibrium" in the process. Points R and R' are alternative equilibria. At R', marginal profits are negative. At R, they are positive. Industry B shifts discontinously from R' to R because at this point its demand curve has moved far enough upward to make R available. At R, industry B is receiving high enough marginal profits from sales of commodity B to pay the bribe and break even in the long run.

The above result, however, is sensitive to the shape of the curve $\frac{\partial \pi_B}{\partial X_B}$ versus X_P. If this curve is less bowed than $\frac{\partial \pi_A}{\partial X_A}$ versus X_P, equilibrium will be at either $X_P = Z$ or at $X_P = 0$. In the former case, no bribery is paid; industry B is absorbing pollution and earning high enough marginal profits to stand the real loss and break even in total profits. In the latter case, industry B buys off all pollution it would otherwise have received from A. The relevant quantity X_P received by B depends on the size (level of activity) in B as well as A. Pollution is a private good in the present context. This does not change the analysis but means that the scale on the horizontal axis in Figure 3 (and other figures) should be changed as the size of industry B (or A) changes.

To this point, only the problems of independent adjustment in A and B have been considered. Interdependent adjustment is complicated by uncertainty about the level of Z from which withholding is measured. Knowledge and agreement on this level would seem unlikely in a dynamic context.[15]

☐ **Conclusions.** It is possible that a stable total profits equilibrium of bribery might be achieved in a dynamic context; but the requirements, as described above, make it implausible. These requirements include: agreement between A and B on the value of Z from which withholding is to take place; demand and supply curves for withholding that have the right relative bow to permit a tangency solution as in Figure 3; and some part of B's bribery contraction curve with positive total profits, and these profits high enough to pay the bribe. This last condition depends on the relative sizes of industries A and B. Size affects the relative prices for conventional commodities A

---

[15] Kamien et al. [9] discuss a different kind of dynamic adjustment. They do not consider the problems discussed above, but analyze the case where industry A might adjust due to such normal causes as increasing demand for output A. Their work conforms to tradition in assuming a single valued function $X_P = f(X_A)$ with X_A independently produced. They consider the case where Z is not known but bribery is based instead on the quantity of X_P that is produced and removed by treatment. With the decision rule for bribery based on the amount treated rather than the amount withheld, they find both X_A and X_P produced in greater amounts than with compensation.

and B and also the internalization price P_W. Finally, there is the conflict of marginal and total profit conditions in A. B may find it possible, through contraction and increases in marginal profits in other lines, to pay the bribe. But A will be receiving a lump sum transfer beyond returns at zero marginal profits, which can only be offset by expanding to negative marginal profits levels elsewhere as long as the bribe is received and total profits are above normal.

Do these results depend on the reference base from which we started, viz., zero profits in industry B in the absence of any pollution? Such a base is entirely proper for the linear homogeneous production functions assumed. But let us relax the assumption of linear homogeneity. What are the conditions that would keep total profits at the zero level in equations (1b) and (3b) with K, S, and Z also as variables? The obvious answer is that somehow the effect of the lump sum transfer must be offset, or approximately offset, by the other transactions in these industries when marginal conditions are satisfied. Strict linear homogeneity may not characterize very many production functions, but it probably becomes closer than would a pair of lump sum production functions geared to produce deficits in A and surpluses in B so that they could be offset by bribery transactions.

It is interesting to note that difficulties of a qualitatively similar sort are encountered in marginal cost pricing for increasing returns industries, though in the increasing returns case, the problem arises from physical production conditions. In the bribery case, it arises from a lack of correspondence between physical and financial conditions. Bribery would match a linear homogeneous physical production function with financial relationships that cannot lead to exhaustion of product.

Compensation, in contrast, creates no such distortions but produces conventional market adjustments and stable equilibria.

5. Some related issues

■ **Blackmail.** Professor Coase expressed concern that if compensation were paid, the damaged party would intensify his activities so as to justify the award of greater damages.[16] The same fear was expressed by Kneese in an earlier work,[17] in the situation where a member of industry B threatens to locate downstream of a firm in industry A. Either way, the latter is faced with making more compensation payments.

The fear is groundless in a competitive setting. The Coase version has our industry B shifting toward more than optimal absorption of waste, which, of course, is in conflict with economic rationality. The Kneese version can be resolved by asking whether B is moving toward or away from optimal adjustment. If B could make a profit with the competitive level of costs and prices, including compensation, then B's locating at that point is socially desirable. It is then A whose operations are subject to question from the standpoint of a welfare maximum. If not, then A has nothing to fear.

Needless to say, a competitive solution is not to be expected in all cases. Market structures are not perfect; nor is knowledge a free good. Bilateral monopoly and/or imperfect knowledge can make blackmail possible. But blackmail then becomes, in the relevant sense, another name for market imperfection and is not unique to the process of internalizing through compensation. Those who believe bribery to be viable could equally well refer to blackmail possibilities for this case and, in fact, Kneese and Bower use the blackmail argument as a reason for recommending against bribery.[18]

---

[16] See [5], pp. 32–33.

[17] See [10], p. 58.

☐ **Third party pricing.** Third party pricing refers to the institutional arrangement where a third party, presumably a Public Authority, sets pollution charges and supervises the conduct of any internalizing transactions. In many contexts, the Public Authority also conducts waste treatment.

Kneese and Bower recommend that the Public Authority collect a tax but pay no subsidy, on the ground that completing the transaction would "lead to inefficient longer-term adjustments."[19] They state, citing Boyd:

> If . . . the right to the use of an asset (or resource) is made contingent upon engaging in a particular activity, and the right to the use of the asset given free to parties engaging in that activity, excessive activity in that line will be generated.[20]

Compensation and bribery are viewed symmetrically, as by Bramhall and Mills.[21] There is the additional twist that the right to use or not use the waterway for waste removal is treated as an asset. This point of view is made clearer in another passage by Boyd.

> The economic maximization model presented here treats the river basin as a multiple-product natural asset. The products of this asset are flows of two types of service, a vector of *waste removal* services and one of *water quality* services. The quantity of the latter services available at a given location depends on quantities of the former consumed at upstream points.[22]

The implied course of action is to charge a rent to both the waste disposer and the quality user. This treats the river facilities as a common asset the use of which should be rationed. It can easily be shown that rationing is quite a different process from internalization. The difference is in the sign that precedes the terms in the objective functions (1) and (3). Labor and capital inputs are being rationed to both industry A and industry B. Well-known first-order conditions require that the marginal value products of each be the same in all uses and equal to the wage rate and interest rate, respectively. In contrast, pollution effects enter (1) and (3) with opposite signs. A and B are on opposite sides of the market in which the pollution externality is to be internalized.

One of the implications of the foregoing is that ownership must be functionally identified with the commodity to be rationed (or discommodity to be disrationed). Decentralized decision-making is possible if property "rights" (responsibilities) are forced on industry A so that if pollutants produced by A were received by B, compensation would be due the latter and B would become "owner." That this is not institutionally impossible is suggested by the contemporary existence of waste removal firms and enforced ownership of wastes, particularly of solid wastes in the form of trash, where payment for removal takes the form of compensation to the remover. In the river basin case, we found that the true rationing problem was not of river capacity but of downstream removal capacity.

---

[18] In [11], pp. 104–5.

[19] *Ibid.*, p. 100. The word "subsidy" is theirs. It has inappropriate connotations for an internalization payment.

[20] *Ibid.*, p. 87.

[21] In [3].

[22] [2], pp. 199–200. Italics are in the original.

☐ **Public goods.** The logic of private goods analysis is equally applicable to public goods such as aesthetic and recreational damages. As long as partial equilibrium models are used, the optimum level of pollution and the compensation price can be determined from the vertical summation of (positively sloping) demand curves of damaged public goods recipients.

An unexpected windfall results from the reversal of the transaction, as required by compensation. In contrast to the situation with positive public goods, where the prospect of paying one's demand price gives an incentive to conceal demand, there is with negative public goods and compensation the incentive to reveal demand.The Public Authority is likely to have ample information volunteered from damaged parties (and from some not damaged). Its problem would be more that of establishing authenticity and accuracy.

The last is a form of the blackmail problem. One way to deal with it is to make compensation in the form of additional treatment to remove more pollution. Instead of paying compensation in financial terms, the control of blackmail might force the payment of compensation in kind. Only authentic damage receivers will benefit from having additional pollution removed.

It is obvious that payment in kind by the above system introduces inefficiencies. By the definition of optimum treatment as the supply-demand equilibrium, there will be nonoptimal (excessive) treatment. The reader is left to catalog other sources of inefficiency. Nevertheless, compensation gives us another approach to the problem of public goods pricing (albeit only for negative public goods). We are probably no further from an optimal translation of theory into practice with compensation than we are with conventional public goods pricing.

Glossary of notations

■ Variables

C	Quantity of capital.
L	Quantity of labor.
P	Price.
r	Return on capital.
w	Wage rate.
X	Quantity of commodity.
λ	Lagrange multiplier.
π	Profits.

☐ Subscripts and superscripts

A	Industry or commodity.
B	Industry or commodity.

J	All-purpose subscript that can take on meaning of N, P, or W.
N	Nectar.
P	Pollutant.
W	Withheld pollutant.

References

1. Bator, F. M. "The Anatomy of Market Failure." *Quarterly Journal of Economics*, Vol. 72, No. 288 (August 1958), pp. 351–79.
2. Boyd, J. H. "Collective Facilities in Water Quality Management," in A. V. Kneese and B. T. Bower, *Managing Water Quality: Economics, Technology and Institutions*, Baltimore, Md.: Johns Hopkins Press, 1968, Appendix to Chapter 10.
3. Bramhall, D. F. and Mills, E. S. "A Note on the Asymmetry Between Fees and Bribes." *Water Resources Research*, Vol. 2, No. 3 (July–September 1966), pp. 615–16.
4. Buchanan, J. M. and Stubblebine, W. C. "Externality." *Economica*, Vol. 29, No. 116 (November 1962), pp. 371–84.
5. Coase, R. H. "The Problem of Social Cost." *The Journal of Law and Economics*, Vol. 3, No. 2 (October 1960), pp. 1–44.
6. Davis, O. A. and Whinston, A. B. "Externalities, Welfare and the Theory of Games." *Journal of Political Economy*, Vol. 70, No. 3 (June 1962), pp. 241–62.
7. ———. "Some Notes on Equating Private and Social Cost." *Southern Economic Journal*, Vol. 32, No. 2 (October 1965), pp. 113–26.
8. Dolbear, F. T., Jr. "On the Theory of Optimum Externality." *The American Economic Review*, Vol. 57, No. 1 (March 1967), pp. 90–103.
9. Kamien, M. J., Schwartz, N. J., and Dolbear, F. T., Jr. "Asymmetry Between Bribes and Charges." *Water Resources Research*, Vol. 2, No. 1 (January–March 1966), pp. 147–57.
10. Kneese, A. V. *The Economics of Regional Water Quality Management*. Baltimore, Md.: Johns Hopkins Press, 1964.
11. ———, and Bower, B. T. *Managing Water Quality: Economics, Technology, and Institutions*. Baltimore, Md.: Johns Hopkins Press, 1968.
12. Kuenne, R. E. *The Theory of General Economic Equilibrium*. Princeton, N. J.: Princeton Univ. Press, 1963.
13. Meade, J. E. "External Economies and Diseconomies in a Competitive Situation." *Economic Journal*, Vol. 62, No. 245 (March 1952), pp. 54–67.
14. Turvey, R. "On Divergences Between Social Cost and Private Cost." *Economica*, Vol. 30, No. 119 (August 1963), pp. 309–13.
15. Wellisz, S. "On External Diseconomies and the Government-Assisted Invisible Hand." *Economica*, Vol. 31, No. 124 (November 1964), pp. 345–62.

Models for determining least-cost investments in electricity supply

Dennis Anderson

Economics Department
Transportation and Public Utilities Division
International Bank for Reconstruction and Development

This paper reviews models used in the electricity supply industry for appraising investments, and presents some extensions. Quantities demanded and the prices of inputs and outputs are assumed to be exogenous, and the models search for investments having the lowest costs. Optimization is over several time periods. Typical decision variables considered are: choice of fossil, nuclear, single- or multi-purpose hydro plant; locations of plants; directions of electrical energy transport (interconnection); timing of investments; replacement; and in all cases the optimum mode of system operation (including hydro storage policy). These variables may be analyzed by linear, non-linear, and dynamic programming as well as other methods. Both global models and optimization treatment of subproblems are reviewed.

1. Introduction

■ In postwar years, the electric power industries of many high- and low-income countries have expanded at average rates of 7 percent per year to as much as 20 percent per year, requiring investments of the order of $150 billion in the U. S. and $1.5 billion in a developing country the size of Colombia; and it is expected that total investments will exceed such magnitudes in the next decade. The problems of determining optimum investment policies in the face of such rapid increases of demand, high costs, the large number and diversity of alternative investment policies, and the numerical tedium of evaluating in depth even a single policy have motivated the development of mathematical models to assist the engineer in scanning and costing alternative policies. This paper reviews these models and presents some extensions to the linear programming (LP) versions.

The author trained in electrical and mechanical engineering and physics at Imperial College and Manchester University and, later, in econometrics at the London School of Economics. He worked for a number of years as an engineer and physicist for the Central Electricity Generating Board, United Kingdom (U. K.), where he also served an apprenticeship, and on industrial project analysis in the former Ministry of Technology, U. K. Mr. Anderson is now working on public utility economics for the International Bank for Reconstruction and Development (IBRD).

This paper is a condensation of a study for the IBRD [2]. Its contents are for the most part derived from the work of others and from working with and talking to others in this field. Particular thanks are due to Narong Thananart of the IBRD for writing many programs to test and apply the various models discussed, and for correcting some mistakes in the formulations in Section 6. The following people have been most generous in communicating their ideas and experience to the

The investment decision variables of the industry interact strongly at a point in time and over time. This occurs for a number of reasons, which are perhaps most easily explained through two examples. First, different energy sources have complementary functions in modern interconnected power systems. The main sources are single- and multi-purpose hydro schemes, of widely varying power and energy storage capacities;[1] fossil fuels, mainly fuel oil, coal, gas, and lignite; thermal and (eventually) fast neutron breeder reactors; and special-purpose peaking plant, mainly gas turbines and pumped storage. Gas turbines have low capital but high generation costs; fossil, higher capital but lower generation costs; nuclear, high capital and low generation costs; and hydro, high or low capital costs (depending on the site) and near-zero generation costs, but with constraints on energy output which may stem from the multi-purpose nature, water inflows, or both. Gas turbines are thus used for peak loads; fossil, for loads of longer duration; nuclear, for base (continuous) loads; and hydro, somewhere in between, depending on the energy constraint. The optimum balance of plant in the system at any point in time will depend on the relative capital and generation costs of the alternative energy sources.

Second, the optimum balance will depend on both the inherited and the expected structure of the power system. For example, more nuclear and less fossil in future years means that the future system fuel savings of hydro schemes installed now will be less; a large nuclear power program in future years may thus shift the present balance towards more fossil and less hydro. Similarly, if the inherited structure is predominantly fossil, then the present emphasis will be on more nuclear and/or hydro to save on system fuel costs.

Because of these kinds of interaction among decision variables, models must be multi-dimensional and couched, as Turvey has said, in terms of historical dynamics. The investment decisions to be taken at the present time depend upon the past and future evolution of investments and thus upon the past and future evolution of factor prices. We shall find that the models discussed below are designed to capture this problem.

Although developed by engineers and operations researchers in the industry, and specifically concerned with investment decisions, these models are not without interest to economists. They have been

---

present writer: Frank Jenkin, Ivan Whitting, George Hext, Eric Parker, and Bill Billington, on work in the U. K.; M. Stengel and M. Pouget of Electricité de France; Mr. Askerlund and his colleagues of the Statens Vattenfallsverk, Sweden; Mr. van der Tak, Mr. Berrie, and Mr. Russell of the IBRD; Mr. John Rixie of AID and his colleagues in the American Institute of Electrical and Electronics Engineers (IEEE); and Professor Alan Manne and Dr. Ralph Turvey. Professor Paul MacAvoy and an anonymous reviewer also provided very helpful comments on an earlier draft. All views, mistakes, misinterpretations, etc., are of course those of the present writer.

The reference list following the text contains papers cited in the text plus other papers which also contain material of relevance for this subject. The initials PSCC refer to the Power Systems Computation Conferences organized by Queen Mary College, London. See note 55.

[1] In Turkey, for example, where we are currently applying LP models, there are over 168 as yet untapped hydro sites, ranging in size from 1 MW to a proposed 4,000 MW multi-dam complex on the River Euphrates, capable of irrigating 700,000 hectares.

occasionally applied in the cost-benefit studies of single- and multi-purpose hydro schemes.[2]

A second and increasingly common application is the determination of the marginal cost structure of the industry for purposes of pricing policy. Bessière and Massé and Turvey in particular have demonstrated the practical value of linear and non-linear programming (LP and non-LP) global models for this purpose,[3] and the pricing models of Littlechild, Pressman, and Williamson, for example, can be viewed as approximations of the global models discussed below.[4]

A third, but as yet unexplored, aspect of these models is their relation to the empirical studies of the type undertaken by Nerlove, Johnston, Galatin, and Dhrymes and Kurtz.[5] These workers have attempted to estimate economies of scale and technical progress in the industry using Cobb-Douglas and CES types of empirical relations between the factor inputs and an optimizing condition for a single time period. The cost structure and the relationships among the factor inputs in the electricity supply industry are, however, defined precisely in the engineers' models without recourse to such empirical relationships. Moreover, it is the daily occupation of planning engineers in the industry to search for optimum investment programs over many time periods. Taken into consideration are economies of scale attainable from large units, external economies of scale attainable from interconnection, technical progress embodied in new equipment, substitution among factor inputs, replacement, the putty-clay nature of the investment decision, the putty-putty nature of the operating decision, the possibility of storage (hydro schemes, pumped storage), and as noted above, the past and expected future evolution of the system. It would seem therefore that the engineers' models are not without significance for econometricians who wish to study the industry.[6]

The plan of this paper is as follows. We begin in Section 2 by formulating the investment problem in cost minimization form. We then review the various approaches used to find optimum solutions. These are three classes which we review in Sections 3, 4, and 5, respectively: marginal analysis, marginal analysis using simulation models, and global models. We shall find that while they are outwardly different in form—ranging from graphical devices and marginal analysis to dynamic, linear, and non-linear programming—this is a difference only of algorithms; they are different methods of solving the same kind of problem. We shall also find that they are often complementary approaches in the following sense. Global models can only give approximate answers in most practical situations. The reason is that the practical details of the alternative programs, particularly of the individual projects in the programs, are too numerous to be handled in one computer run. Having obtained approximate solutions from the global models, we then turn to marginal analysis using simulation models to focus on the fine details of individual project selection and design. Finally, in Section 6, we present three LP extensions to the global and simulation models reviewed. They cover (1) a fresh treatment of replacement, (2) the introduction of decision variables for hydro storage capacity and storage policy, and (3) regional decision variables, to give a fuller treatment of transmission.

---

[2] See, for example, the study of water resource systems by Maass et al. [55], the IBRD study of the Indus Basin Project by Jacoby [43] and Lieftinck et al. [51], the simulation studies of Hufschmidt and Fiering [42], and the discussions of Turvey [76], Prest and Turvey [71], Eckstein [23], and Krutilla and Eckstein [48]. More recently, Forster and Whitting [29,30] and Fernandez and Manne [27,28,58] have been linking these models into wider studies of resource allocation in the energy sector and the economy.

[3] See [9,10], and [77,79], respectively. See also Dr. Turvey's note in the *Economic Journal* [80].

[4] See [54], [70], and [86], respectively.

[5] See [66], [44], [32], and [21], respectively.

[6] Turvey makes a similar point, but more strongly. See [78], p. 8.

Before proceeding, let us make clear a number of limitations of this paper:

(1) The quantities demanded are assumed to be exogenous, and the objective is always cost minimization. This assumption could be relaxed if required so as to maximize consumers' plus producers' surplus. The papers of Littlechild and Pressman[7] would be good starting points in this respect. It is thought, however, that the most practical way to treat interactions of demand and supply when formulating an investment program is by iteration, taking demand as given (but hopefully related to some kind of rational pricing policy), searching for least-cost solutions, and then revising demand estimates on the basis of marginal costs and prices.[8] In connection with formulating a pricing policy, Turvey has also argued for an iterative approach.[9]

(2) Use of one or more investment models is the first of several stages of the investment decision process. Engineering analysis of solutions follows and generally requires a revision of the solutions. The investment program finally selected must satisfy a number of engineering criteria regarding system stability, short-circuit performance, the control of watts, vars, and voltage, and the reserves and reliability of supply.[10] The search for an investment program which satisfies engineering and economic criteria is an iterative, multi-disciplinary process.

(3) All the formulations presented are deterministic. Allowances are of course made for uncertainties in demand, plant availability, and flows of water to hydro schemes, but in the simple form of margins of spare capacity. This is frequent practice, although people are working with stochastic counterparts to the models presented and their work is noted.

(4) There is no discussion of terminal conditions as analyzed by Hopkins[11] or of the optimum breakdown of the time period of the study into discrete periods.

(5) There is no discussion of the dual variables from the LP models or of pricing policy. We thus neglect much important work of Bessière and Petcu, Turvey, Littlechild, Williamson, and many others.[12]

---

[7] See [54] and]70].

[8] See Berrie [6].

[9] See [79], p. 288.

[10] See Stagg and El-Abiad, for example [74].

[11] In [41].

[12] See [11], [77,79], [54], and [86].

(6) Finally, we do not explore the connection between these models and those customarily used in econometric research (i.e., of the type mentioned above).[13]

2. The investment problem in cost minimization form

■ The principles of the following formulation were first enunciated in the early 1950s by Massé and Gibrat, who solved the problem using linear programming.[14] A subsequent paper by Bessière and Massé and the book by Massé formulate the problem more generally.[15]

The search for an optimum (least-cost) investment program also entails, for each plant program considered, the search for an optimum operating schedule. Let the power capacity of any plant in the system be defined by X_{jv}, j denoting the type of plant (hydro, fossil, nuclear, etc.) and v the vintage (year of commissioning). Also, let the power output of this plant at any instant t be $U_{jv}(t)$, $0 \leq U_{jv}(t) \leq X_{jv}$. The operating costs of this plant over the interval $t = 0$ to T are given by:

$$\int_{t=0}^{t=T} F_{jv}(t)\cdot U_{jv}(t)\cdot dt\,,$$

where $F_{jv}(t)$ are the discounted operating costs per unit of energy output.

At any instant t the operator has before him $j = 1\cdots J$ types of plant of different vintages, comprising the initial plant composition of the system, $v = -V$ to 0, and the plant installed (at discrete intervals) between 0 and t. To obtain the total system operating costs in the interval dt, we must summate over all vintages $v = -V$ to t and over all types of plant. The total future operating costs are then

$$\int_{t=0}^{t=T} \sum_{v=-V}^{t} \sum_{j=1}^{J} F_{jv}(t)\cdot U_{jv}(t)\cdot dt\,.$$

The investor's objective is to minimize the sum of capital and operating costs over some future time period 0 to T:

$$\text{Minimize} \sum_{v=1}^{T} \sum_{j=1}^{J} C_{jv}\cdot X_{jv} + \int_{t=0}^{T} \sum_{v=-V}^{t} \sum_{j=1}^{J} F_{jv}(t)\cdot U_{jv}(t)\cdot dt\,, \quad (1)$$

where C_{jv} are the capital costs per unit of capacity of plant j, vintage v. All costs of course be expressed as social opportunity costs.

The discrete approximation to (1) is often a more convenient function to use:

$$\text{Minimize} \sum_{v=1}^{T} \sum_{j=1}^{J} C_{jv}\cdot X_{jv} + \sum_{t=1}^{T} \sum_{v=-V}^{t} \sum_{j=1}^{J} F_{jvt}\cdot U_{jvt}\cdot \theta_t\,, \quad (2)$$

where θ_t is the width of the time interval considered at time t.

---

[13] Page 269 *infra*.

[14] See Massé and Gibrat [63] for an English translation of their original paper or *Management Science*, Vol. 3, No. 2 (January 1957), pp. 149–66.

[15] See Bessière and Massé [10] and Massé [62]. Bessière [9] outlines the development of methods used by Electricité de France during the 1950s and 1960s.

The search for optimum capacities (optimum X_{jv}) and for optimum operating schedules (optimum U_{jvt}) is of course subject to a number of important conditions. First, sufficient plant must be operating at all times to meet the instantaneous power demand, which we define by Q_t. Thus

$$\sum_{j=1}^{J} \sum_{v=-V}^{t} U_{jvt} \geq Q_t , \qquad t = 1 \cdots T . \tag{3}$$

Only at times of peak load will all of the plant be in operation.[16] At other times the instantaneous power demand can be met with much of the plant not operating (i.e., some of the U_{jvt} are zero); this plant will in general be that with the highest operating costs.

Second, no unit of plant can be operated above its peak available capacity:

$$0 \leq U_{jvt} \leq a_{jv} \cdot X_{jv}, \qquad \begin{array}{l} j = 1 \cdots J \\ v = -V \cdots t \\ t = 1 \cdots T, \end{array} \tag{4}$$

where a_{jv} is the availability of plant j, vintage v (a_{jv} is usually about 0.9).[17] Note that X_{jv} for $v = -V$ to 0 are predefined constants and represent the capacities of the inherited capital stock.

Third, there may be constraints on the operation of hydro plant. Seasonal shortages of water inflow, or the requirements of irrigation and flood control, will impose restrictions on the amount of electricity to be generated in, e.g., a given season. There will still be a choice however on the timing of the hydro operation within the season (in general, it will be operated at times of peak demand, when fossil energy is most expensive). The simplest form the hydro constraints take is the following. Let H_{vs} be the hydro-electric energy to be delivered in season s by the hydro scheme of vintage v. Then we must choose the decision variables for hydro operation, $U_{jv}(t)$ (j = hydro = h) such that, in minimizing total system operating costs, all available hydro energy will be utilized:

$$\int_{\text{season } s} U_{hv}(t) \cdot dt = H_{vs} . \tag{5}$$

(At a later stage in this paper we shall be discussing methods of searching for optimum values of H_{vs}.)

Fourth, there are constraints to represent what the French writers call the "guarantee conditions." These are to guarantee supply, to an unacceptable probability limit, in event of contingencies—water shortages in dry seasons, peak demand above mean expectations, or plant outage. These constraints take two forms, one to guarantee peak power supplies, and the other to guarantee energy supplies in critical periods. Let us take them in turn. Following Massé and Morlat,[18] suppose that e is the probability that the yearly peak demand, defined by $\hat{Q}_t$, will be met; that is, the probability that the aggregate available capacity is greater than $\hat{Q}_t$ is e, as follows:

$$Pr\left(\sum_{j=1}^{J} \sum_{v=-V}^{t} a_{jv} \cdot X_{jv} - \hat{Q}_t \geq 0\right) = e, \quad t = 1 \cdots T , \tag{6'}$$

---

[16] An allowance for plant outages will be considered shortly.

[17] Data on a_{jv} for different countries are discussed by Cash and Scott [17].

[18] See [60] and [64], respectively.

where a_{jv} and $\hat{Q}_t$ are stochastic variables. This is a "chance constraint" of the type discussed by Charnes and Cooper.[19]

When reviewing the practices adopted by European countries in planning system security, Cash and Scott found that the actual choice of e varies widely between countries.[20] They also found that, while the choice is sometimes backed up by statistical and economic calculation, it is generally determined by experience. This is partly because estimating the costs and benefits to the economy of a given level of security is exceedingly difficult and subject to large errors, and partly because the required level may rest on a number of noneconomic factors—e.g., the hostility of press and public opinion in event of supply shortages (which, almost by definition, seem to occur when least desired). Finally, they found that most countries think of the guarantee condition in terms of a "margin of available capacity" over and above what is required to meet the mean expected peak demand.

The guarantee condition (6′) is therefore frequently simplified in practice. Let m be the margin of spare available capacity required to meet demands above the mean expectation; then (6′) is expressed as

$$\sum_{j=1}^{J} \sum_{v=-V}^{t} a_{jv}\cdot X_{jv} \geq \hat{Q}_t(1+m), \quad t = 1\cdots T, \tag{6}$$

where a_{jv} and $\hat{Q}_t$ are once again mean expected quantities.

It should be added that while condition (6) is much simpler than condition (6′), it by no means implies a loss of rigor in the planning process. On the contrary, calculating the probability distribution of available capacity in various regions of a modern interconnected system is itself a highly complex and specialized computation which may require the use of Monte Carlo techniques. For this reason it is perhaps best treated as a separate calculation, even though this may result in repetition and modification of the least-cost exercises. System planning, as we remarked in the introduction, is an iterative, multi-disciplinary operation.

A similar pattern of discussion follows when we examine the problem of guaranteeing energy supplies in dry seasons on mixed hydro-thermal systems. Let e now denote the probability that the potential energy available from both hydro and thermal plant will be greater than the energy demand in the critical period; and let $t = t'\cdots t''$ represent the critical period. The potential energy output from a thermal plant is limited by its available capacity, $a_{jv}\cdot X_{jv}$. The potential energy output from a hydro plant is limited by constraints of type (5). The guarantee condition is therefore

$$Pr\{\sum_{t=t'}^{t''} (\underset{(j \neq \text{hydro})}{\sum_{v=-V}^{t} \sum_{j=1}^{J} a_{jv}\cdot X_{jv}} + \underset{(j \neq \text{thermal})}{\sum_{v=-V}^{t} \sum_{j=1}^{J} U_{jvt}} - Q_t)\theta_t \geq 0\} = e\,, \tag{7′}$$

where a_{jv}, U_{jvt}, and Q_t are stochastic.

---

[19] See "Chance-Constrained Programming," *Management Science*, Vol. 6, No. 1 (October 1959), pp. 73–79, "Chance Constraints and Normal Deviates," *Journal of the American Statistical Association*, Vol. 57, No. 297 (March 1962), pp. 134–48; and "Deterministic Equivalents for Optimising and Satisficing Under Chance Constraints," *Operations Research*, Vol. 11, No. 1 (January/February 1963), pp. 18–39.

[20] See [17].

Again it is possible to approximate the chance constraint (7′), postponing a more rigorous study of the reserves and reliability of supply for a separate stage of the planning process. Let β_{jv} (with $j =$ hydro) to be the ratio of the energy output of hydro plant j, v, in the critical period of a dry year, to its mean expected output in this period of an average year. Then (7′) simplifies to

$$\sum_{t=t'}^{t''} \sum_{v=-V}^{t} \Big(\sum_{\substack{j=j \\ (j \neq \text{hydro})}}^{J} a_{jv} \cdot X_{jv} \cdot \theta_t + \sum_{\substack{j=1 \\ (j \neq \text{thermal})}}^{J} \beta_{jv} \cdot U_{jvt} \cdot \theta_t\Big) \geq \sum_{t=t'}^{t''} Q_t \cdot \theta_t \,, \tag{7}$$

where a_{jv}, U_{jvt}, and Q_t are once again mean expected quantities. Note that for hydro schemes with highly uncertain water supplies, β_{jv} will be low, and if such schemes are included, then other thermal or hydro capacity will be required to satisfy this constraint. Thus β_{jv} indirectly applies a cost penalty on schemes according to the variability of their supplies.

To complete the formulation of the investment problem, there will generally be a number of "local" constraints. Examples are constraints to limit expenditures on capital or foreign exchange, regional development constraints (e.g., a lower limit to the use of coal or hydro resources), political and social constraints, etc.

It will be evident that this formulation can be extended or contracted in many ways. Simulation models, for example, extend it in one direction and contract it in another. Essentially the X's (i.e., the investment plans) are predefined constants in a simulation model; this leaves more computer space (core storage) available to examine the U's (the dispatching schedules) in more detail. Another extension is to treat uncertainties in costs, water supplies, and demand forecasts through stochastic programming.[21] Some formulations have also been adapted to optimize the operation of multi-purpose, multi-dam hydro schemes.[22] Replacement decisions have been included in the objective function.[23] Efforts are being made to treat both supply and demand on a regional basis.[24] Finally, some workers are embodying the formulation for the electricity sector into larger models of the energy sector and of the economy.[25] We shall discuss some of these and other extensions subsequently.

3. Marginal analysis

■ Marginal analysis was first applied to investments in electricity supply by Electricité de France in the late 1940s.[26] Since then it has been applied regularly in many other countries.[27] The analysis starts from an arbitrary but reasonable initial program—a "reference solution"—and then seeks to improve it (reduce costs) by marginal substitutions. The reference solution and the solution obtained after a marginal substitution has been made satisfy the same power and energy demands. Whenever the cost function is convex,[28] marginal analysis should ultimately lead to a uniquely optimum investment and operating program over time.

---

[21] E.g., Bessière [8,9], Electricité de France [25], Fernandez and Manne [27], Lindqvist [52], Little [53], and Gessford and Earlin [34] all discuss models which have some stochastic elements, either for demand or water supply variables.

[22] E.g., Jacoby [43] and United Nations Economic Commission for Europe [83].

[23] See Massé [62].

[24] See Electricité de France [25]; Fernandez, Manne, and Valencia [28].

[25] Forster and Whitting [30]; Manne [58].

[26] See Giguet [35], Massé [60], and also Bessière's note [9] and the bibliography contained therein.

[27] E.g., United Nations [82] and van der Tak [85].

A common application of marginal analysis has been the comparison of fossil and hydro alternatives to meet a given demand for electricity. The hydro plant may require a higher investment (I) than fossil ($I_h > I_f$), but the total system operating costs in subsequent years are less. The total, discounted system operating costs at time t are

$$(1 + r)^{-t} \sum_{j=1}^{J} \sum_{v=-V}^{t} F_{jvt} \cdot U_{jvt} \cdot \theta_t = (1 + r)^{-t} \cdot \phi_{th}$$

if the hydro project is adopted and

$$(1 + r)^{-t} \cdot \phi_{tf}$$

if the fossil project is adopted.

The present worth of the savings if hydro is substituted for fossil is then

$$\mathrm{PW} = (I_f - I_h) + \sum_{t=0}^{T} (1 + r)^{-t} (\phi_{tf} - \phi_{th}) \,, \qquad (8)$$

and according as this value is positive or negative the hydro is or is not preferable to the fossil investment. The value of PW savings calculated in (8) is sometimes known as the relative profitability of the hydro investment, since the calculation shows whether or not the hydro investment improves upon the reference solution.

Among the advantages of this calculation are its practical simplicity and the feature that it is easy to adjust the arithmetic for many local costs and benefits of a project. For example, different locations of hydro and fossil stations will lead to different transmission costs; maintenance costs for hydro stations are lower; the fossil station (lifetime about 30 years) will be replaced before the hydro (lifetime about 50 years), so that the discounted replacement costs may have to be included; the hydro may have flood control benefits; the fossil plant may induce more employment in local coal mines; and so on. Such local features can readily be included in the arithmetic.[29] The calculation can also be readily formulated for comparisons between nuclear and fossil plant at base load; or between fossil, pumped storage and gas turbines, at peak load.[30]

There are, however, two difficulties with marginal analysis. First, it is tedious to calculate operating and fuel costs over a 20- or 30-year period, when the demand fluctuates rapidly by the hour, when there may be four or more types of plant on the system, 30 or more vintages of plant, and when the expansion of the system introduces new vintages and types of plant while replacing others. It is also an optimizing problem in itself, since each plant must be located on the system operating schedule[31] so as to minimize total fuel and operating costs. Second, the marginal substitutions to the investment plan may be many, requiring special routines to scan and cost the alternatives. It was to overcome the first difficulty that simulation models were developed, and to overcome the second that global models were developed. We discuss these in turn.

---

[28] Phillips et al. [69] prove convexity for the mixed fossil-nuclear system of the U. K. Their model is described later (Section 5).

[29] See, e.g., van der Tak [85], Boiteux [14], and Turvey [76].

[30] See Openshaw-Taylor and Boal [68], who also consider many other kinds of decision—e.g., on transmission and distribution equipment.

4. Simulation models

■ We discuss three formulations:

(1) Models which integrate the load duration curve directly,
(2) Models which use dynamic programming, and
(3) Models which use linear programming.

Non-linear programming could also be used, but the present writer did not find LP or non-LP simulation models discussed in the literature.[32]

☐ **(1) Direct integration of the load duration curve.**[33] Simulation models which integrate the load duration curve directly are particularly suitable for power systems having thermal plant only,[34] although they have been adapted for mixed fossil-hydro systems by Jacoby.[35] We shall first consider the thermal power system and then look at Jacoby's model.

On an all-thermal system the cheapest way of meeting the demand at any point in time is to run the stations with the lowest operating costs. The system operator tabulates the power stations in ascending order of marginal operating costs and loads and unloads the stations sequentially as the demand rises and falls (merit-order operation). We picture this situation graphically in Figure 1. For clarity we aggregate the system into four representative power stations; in ascending order of marginal operating costs they are nuclear, new fossil, old fossil, and gas turbines. By projecting the plant capacities horizontally through the daily demand curves of frame (b), we see the times when the different plants are started up, loaded, unloaded, and shut down on different days. By continuing the projection horizontally through the load duration curve, frame (c), we can find the total operating time of each plant for the period represented by the curve. By estimating the areas sliced out of the load duration curve, we can estimate the energy delivered by each plant and thus the total system operating costs. These costs will be at a minimum under this arrangement because the plant with the highest operating costs (older vintage of fossil plant and gas turbines) will be operated the least.

This type of simulation model is used by the Central Electricity Generating Board, U. K., for the estimation of generation savings associated with different investment programs. It is reported by Berrie and Whitting and by Jonas.[36] The simulation can be refined

---

[31] Sometimes called the load dispatching schedule.

[32] LP and non-LP models are used extensively for real-time scheduling [see Section 4(3)].

[33] The load duration curve is defined in the Appendix.

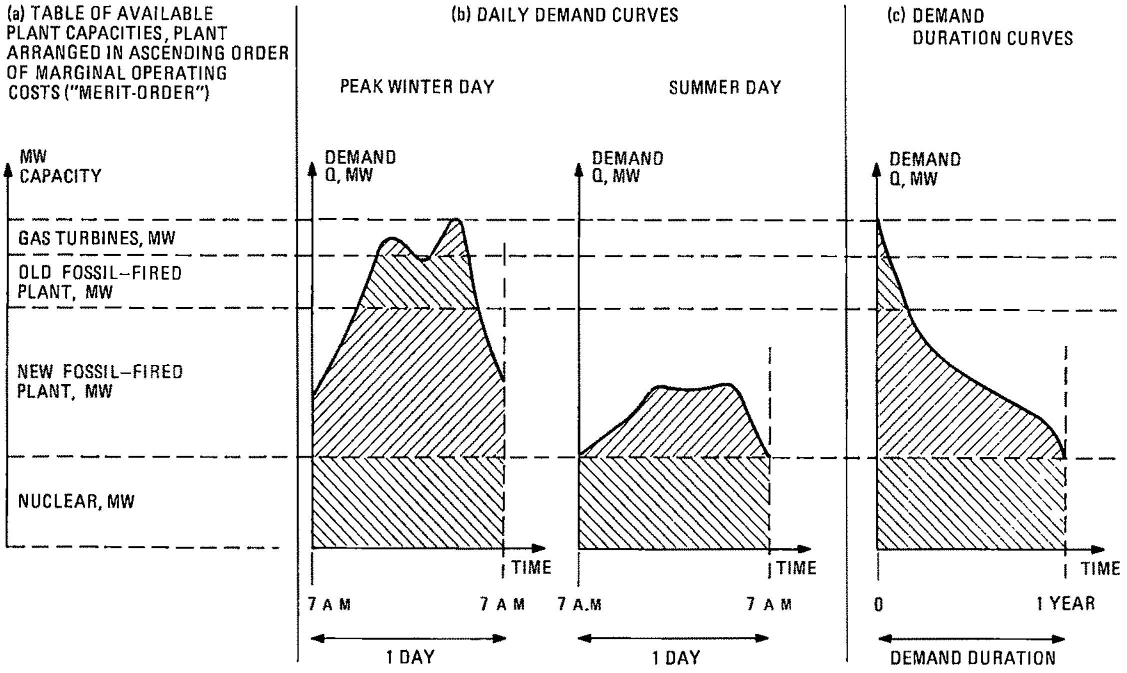
[34] The Central Electricity Generating Board in the U. K. has pursued this method in meticulous detail. See for example Jonas [45], Whitting and Berrie [7], and the review paper of Berrie [6], pp. 22, 29.

[35] In [43]. See also Lieftinck, Sadove, and Creyke [51], Vol. III.

[36] See [5,7] and [45], respectively.

FIGURE 1

LOAD DISPATCHING ON AN ALL-THERMAL SYSTEM


(a) TABLE OF AVAILABLE PLANT CAPACITIES, PLANT ARRANGED IN ASCENDING ORDER OF MARGINAL OPERATING COSTS ("MERIT-ORDER")
MW CAPACITY
GAS TURBINES, MW
OLD FOSSIL-FIRED PLANT, MW
NEW FOSSIL-FIRED PLANT, MW
NUCLEAR, MW
(b) DAILY DEMAND CURVES
PEAK WINTER DAY
SUMMER DAY
DEMAND Q, MW
DEMAND Q, MW
TIME
TIME
7 AM
7 AM
7 A.M
7 AM
1 DAY
1 DAY
(c) DEMAND DURATION CURVES
DEMAND Q, MW
TIME
0
1 YEAR
DEMAND DURATION

in many ways, of which we shall mention three. First, because of maintenance schedules, the table of available plant capacities will differ between seasons. The horizontal projections will only hold, say, for a season, and not for a year as indicated in the diagrams. Thus a different load duration curve is required for each season. Second, the operating costs of each station should be adjusted for transmission losses; this introduces a small quadratic term into the operating cost of each station. Third, the transportation of the coal from several collieries to the power stations is an important element in operating costs. The model determines the operating costs of each power station as follows:

> Trial (marginal operating) costs are used to obtain an initial merit order. A loading simulation study is carried out to obtain trial fuel consumptions, which are then fed into the standard (linear programming) transportation calculation to determine the minimum cost coal allocation and the corresponding station (marginal operating) costs. These are then substituted for the trial value to form a new merit order, this process being repeated until there is no significant change in the costs of generation, when generator loadings and fuel consumptions are consistent with minimum cost fuel allocations.[37]

Monte Carlo studies using this model have also been undertaken to examine the effects on costs of uncertainties in data input, but the results remain unpublished.

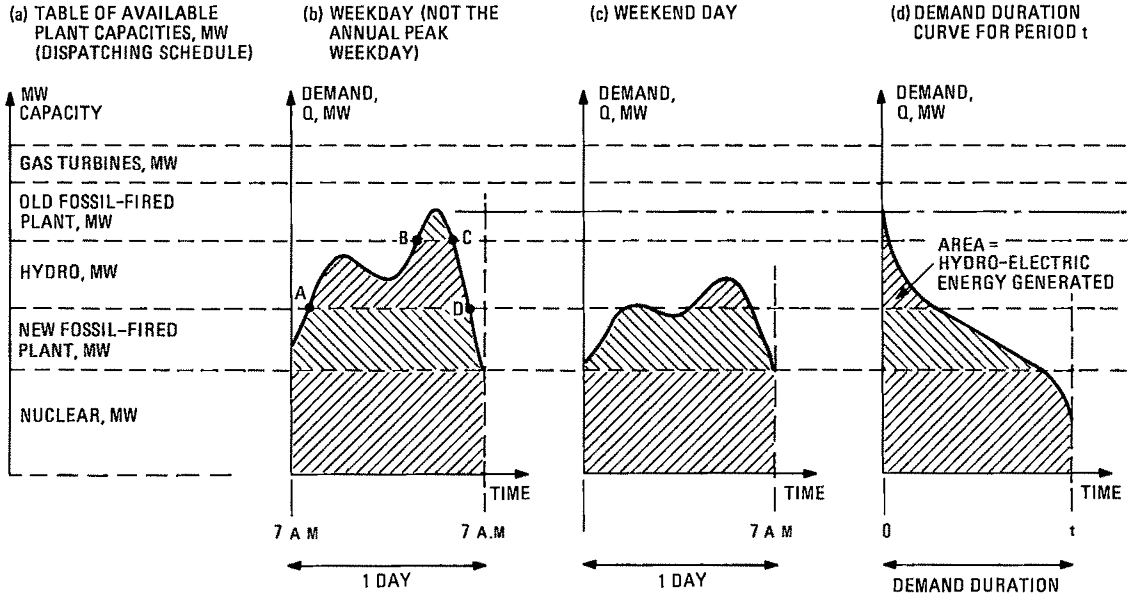
We now look at Jacoby's adaptation of this kind of model for a mixed storage hydro-thermal system. We first consider the optimum position of a single storage hydro station on the power system load dispatching schedule (i.e., the optimum position in the "merit order" table). If a system has several such stations, then the single one we

---

[37] Berrie [5].

FIGURE 2

LOAD DISPATCHING ON A MIXED HYDRO–THERMAL SYSTEM (ONE HYDRO STATION)


(a) TABLE OF AVAILABLE PLANT CAPACITIES, MW (DISPATCHING SCHEDULE)

(b) WEEKDAY (NOT THE ANNUAL PEAK WEEKDAY)

(c) WEEKEND DAY

(d) DEMAND DURATION CURVE FOR PERIOD t

consider is taken to represent the aggregate characteristics of all hydro stations.

Suppose that in any period t the hydro energy allotted for electric power generation is H_t[38] and that the peak power capacity of the hydro is X_{ht}. If the hydro is to maximize fuel savings it must of course discharge the full amount H_t. It must also be operated at times when the fuel costs on the system are most expensive; this happens at times of peak demand (during the period t) when the older and less efficient units of the thermal plant are operating. Consider a particular day [Figure 2(b)]. The hydro plant should begin operating at point A in the morning, generate full power output at B and through to C, and then reduce power to D during the night. However it will not (in the optimum) be delivering full power every day. In the example of Figure 2(c) the hydro is still operated over the peak demand periods of the weekends, but in view of the higher demands and fuel costs during the week, it is cheaper to store energy for delivery during weekdays.

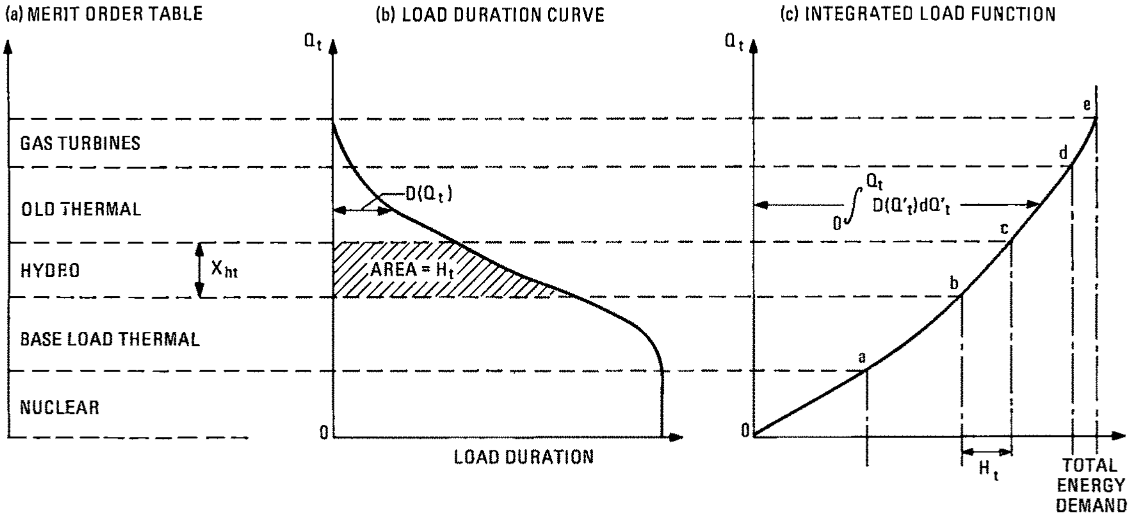
In the optimum, the hydro stations will occupy the same place in the system dispatching schedule every day throughout the period. If they occupied a higher place at the weekends (i.e., were operated at lower values of total system demand) then less energy would be available for operation during the weekdays; the low efficiency thermal stations would then have to supply extra energy during the peak demand periods of the weekdays. If the hydro occupied a lower place at weekends, then extra energy would be available for use during the week, when two things could happen: (1) the hydro would be unable to discharge the extra energy, because of insufficient capacity, and there would be spillover, unless (2) it were to operate at a higher place in the dispatching schedule. If (1) were to happen

---

[38] We assume (as Jacoby does) that H_t is given. Later we consider the problem of finding the optimum value of H_t when there is an option to store hydro energy for the period $t + 1$ [Sections 4(2) and 4(3)].

FIGURE 3

JACOBY'S METHOD FOR DERIVING THE INTEGRATED LOAD FUNCTION



there would simply be a waste of energy. If (2) were to happen then the more efficient thermal plant would be removed from base load operation during the week, and less efficient thermal plant would be operating in place of hydro during the weekends. To conclude: In the optimum the hydro station will occupy the same place in the table of merit order operation every day throughout the period t.

This conclusion enables us to use the load duration curve to determine the optimum position of the hydro station in the merit order table. The horizontal lines which represent the power capacity of the hydro (in Figure 2) must cut the load duration curve at those points (1) where the area cut out of the load duration curve exactly equals the energy to be supplied by the hydro in period t, and (2) where the gap between the lines represents the peak power capacity of the hydro. For two or more hydro plants the technique is the same. Each plant must deliver all the energy allotted for period t, and it must occupy that place on the load dispatching schedule defined above.

The areas under the load duration curve can be calculated graphically or by numerical integration, and the hydro plant located by trial and error. The approach used by Jacoby is, however, much simpler: first to integrate the load duration curve directly and plot the integral (the energy demand) against power demand. This gives us a curve known as the integrated load function shown in Figure 3(c). The energy delivered by each plant can be read directly off the abscissa. The energy delivered by nuclear plant, for example, is obtained by projecting a line vertically downwards into the abscissa from point a. Similarly, projections downwards from points (b,a), (c,b), (d,c) and (e,d) give the energy delivered by base load thermal, hydro, old thermal, and gas turbines, respectively.

These are the types of computation embodied in Jacoby's model, which can estimate operating costs in considerable detail. For example, the total system operating costs may be evaluated for each month of a 20-year period; several hydro units and over 20 to 30 thermal units may be considered. The model was used extensively in the Indus Basin project in Pakistan, and El Chocon in Argentina.[39]

---

[39] See Jacoby [43] and Lieftinck, Sadove, and Creyke [51].

The model can also be formulated to calculate the operating savings associated with transmission links (although this becomes very difficult if there are many regions of generation and demand).

An assumption of the Jacoby model is that H_t, the hydro energy allotted for electric power generation during each month, is known in advance. For large irrigation projects, where power is often a fringe benefit, this assumption is realistic. But it is less realistic in many hydro-electric schemes, when the problem is to determine an optimal water storage policy from period to period. If the next period's demands are high and water inflows low, how much water should be stored for the next period? This type of problem was solved by Massé,[40] who in doing so apparently adumbrated the technique of dynamic programming as applied to this problem. Little[41] formulated the problem explicitly in terms of dynamic programming. Since then the method has been refined and developed by many workers, in particular the Swedish engineers,[42] who use the type of calculations undertaken in Jacoby's model as a routine to compute the total system operating costs for a wide range of storage policies scanned by the dynamic programming (DP) algorithm.[43] It is to this work that we now turn.

☐ **(2) Dynamic programming.** DP techniques have been used by many workers to determine the optimum operating schedules for long-range storage reservoirs (fortnightly, monthly, seasonal reservoirs) on mixed hydro-thermal systems.[44] The method can also be adapted for flood control and irrigation projects.[45] The answers it gives can also be obtained by linear programming.[46] The question to be answered is the following: Given that demand and water supplies fluctuate periodically, how much water should be stored for the next period and how much should be utilized in the present period? The decision process is sequential, since the next period's decisions will depend upon how much water should be stored for the period following that, and so on.

We formulate the problem assuming water supplies are known with certainty. This assumption can be readily relaxed if required.[47] Discrete time intervals are taken. The model below also assumes *one* long-range storage reservoir weighted to correspond to a whole system's reservoirs, one hydroelectric generator (also equivalent to that of the whole system), and a number of thermal stations (fossil and/or nuclear). This follows the practice of all previous writers. It is quite straightforward to extend the model and represent the system's hydro stations with two or more equivalent stations with different storage capacities and water inflow patterns, but this enlarges the dimensions of the problem considerably. Let

---

[40] In [59], Vol. I.

[41] In [53].

[42] See Lindqvist [52].

[43] It is not clear however whether they work with a load duration curve or an integrated load function.

[44] Lindqvist [52] presents the model that was developed for the Swedish State Power Board, which has been used extensively for the technical and economic long-term planning of system extensions. Lindqvist's work builds upon that of Little [53]. See also Koopmans [47] and Gessford and Karlin [34]. The subject of the optimal management of seasonal reservoirs has a long history; e.g., see Massé [59], Vol. I, and Morlat [64].

[45] See for example Manne [57], Thomas and Watermeyer in Maass et al. [55] and Haissman [39].

[46] See Manne [56] and Section 4(3) of this paper.

[47] In the above references (notes 42, 44, and 45), all but Koopmans work with stochastic models.

S_t = storage at the beginning of period t (KWh),
H_t = hydro energy (water) discharge during period t (KWh), and
W_t = water inflow during period t adjusted for losses and expressed in units of potential energy (KWh).

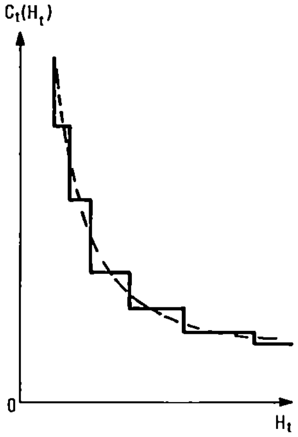
The storage at the end of period t is then

$$S_{t+1} = S_t + W_t - H_t. \tag{9}$$

Suppose we specify a value for H_t. Then following the methods outlined above (e.g., for Jacoby's model), it is possible to determine the optimum system operating schedules and costs for this period. As we increase H_t the total system costs will decrease, because less fuel is burned in the thermal plant. By computing total system operating costs (denoted by C_t) for a range of H_t, we can obtain the kind of curve shown in Figure 4. The shape of this curve (neglecting discontinuities) will generally be concave to the origin because (1) an increase of H_t will always reduce the energy to be delivered by fossil plant, (2) the marginal operating costs of the thermal plant, because of merit-order operation, increase with the amount of thermal plant operated, and (3) the plants with the highest operating costs are generally the oldest and smallest.

FIGURE 4

TOTAL SYSTEM OPERATING COSTS AS FUNCTION OF HYDRO ENERGY UTILIZED IN PERIOD t



The objective is to operate the hydro scheme so as to minimize the total system operating costs over some time period T. That is, we require

$$\underset{H_1 \cdots H_T}{\text{Min}} \sum_{t=1}^{T} C_t(H_t) \tag{10}$$

subject to constraints (9).

This is in fact a standard deterministic inventory problem, which is often solved by the recursive methods of dynamic programming.[48] The principles of the method are as follows. Suppose we fix the amount of water to be stored at the beginning of period t (end of period $t - 1$) at some value of S_t; and suppose also that for this value of S_t we know the values of H_1 to H_{t-1} which minimize total costs up to the beginning of t. We define these costs by ϕ_{t-1}, the minimum of which will depend on the value we have chosen for S_t:

$$\phi_{t-1}(S_t) = \underset{H_1 \cdots H_{t-1}}{\text{Min}} \left\{ \sum_{j=1}^{t-1} C_j(H_j) \right\}. \tag{11}$$

Now suppose that we know $\phi_{t-1}(S_t)$ for a range of values of S_t between zero and, say, $\hat{S}_t$. The next step we can take is to find $\phi_t(S_{t+1})$ for a whole range of S_{t+1} between zero and $\hat{S}_{t+1}$, as follows:

$$\phi_t(S_{t+1}) = \underset{0 \leq S_t \leq \hat{S}_t}{\text{Min}} \left\{ C_t(H_t) + \phi_{t-1}(S_t) \right\}. \tag{12}$$

In view of constraints (9) the hydro flows H_t are implied by S_{t+1} and S_t so that

$$\phi_t(S_{t+1}) = \underset{0 \leq S_t \leq \hat{S}_t}{\text{Min}} \left\{ C_t(S_t - S_{t+1} + W_t) + \phi_{t-1}(S_t) \right\}. \tag{13}$$

---

[48] See for example, Hadley [38], Chaps. 10 and 11.

This can be viewed as a forward recursive decision rule of dynamic programming. For each period t, the system operating costs are evaluated for a range of S_{t+1} and the range S_t considered in the previous scan. For each S_{t+1}, the optimum value of S_t is found from (13). Similarly, at $t + 1$ a range of S_{t+2} is taken, and for each S_{t+2} the optimum S_{t+1} is found; and so on. At the beginning, S_1 is known so that:

$$\phi_1(S_2) = \min_{0 \leq S_1 \leq \hat{S}_1} \{C_1(S_1 - S_2 + W_1)\} . \tag{14}$$

This gives us the starting point. The forward recursions are continued until the optimum decisions of interest are not influenced by extra recursions.

To solve the problem for the case when hydro supplies are treated stochastically, it is necessary to use backward recursive formulas. These formulas (but including the frequency distributions) are a mirror image of the formulas above.

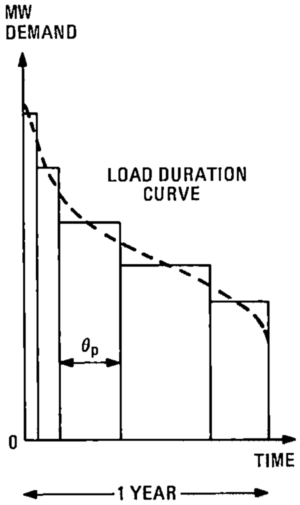
DP simulation models have been applied to many problems. Lindqvist (writing in 1962)[49] informs us that, in Sweden,

> the model has been utilized since the Spring of 1959 for several hundred calculations, e.g., for the calculation of utilization times for nuclear and thermal plants during drought years, for the optimum ratio between hydro-electric and thermal power as a function of interest rates, fuel costs and capital costs; furthermore, for the calculation of the economic consequences of errors in the long-range prediction of net consumption, and for possible secondary deliveries to neighbor countries in the future, etc.

Apparently, the model is still used for purposes such as these.

FIGURE 5

BLOCK REPRESENTATION OF LOAD DURATION CURVE



☐ **(3) Linear programming.** The problem as formulated in Section 2 is already in an LP form. It is convenient to alter the notation slightly and let each period t be represented by a load duration curve broken down into $p = 1 \cdots P$ blocks each of width θ_p (see Figure 5). The periods t may represent months, seasons, or years, etc., according to the approximation desired. If there is to be seasonal or monthly storage hydro on the system, t will accordingly represent seasons or months. Since the capacity variables X_{jv} are predefined constants in the simulation model, the objective is to choose the operating decision variables U_{jtvp} such that the total system operating costs are at a minimum. Thus [see expression (2) above]:

$$\text{Minimize} \sum_{j=1}^{J} \sum_{t=1}^{T} \sum_{v=-V}^{t} \sum_{p=1}^{P} F_{jtvp} \cdot U_{jtvp} \cdot \theta_p . \tag{15}$$

This objective is subject to the capacity constraints

$$0 \leq U_{jtvp} \leq a_{jv} \cdot X_{jv}, \quad \text{all } j, t, v, p \tag{16}$$

and to the constraints that aggregate output must be sufficient to meet the demand at all times:

$$\sum_{v=-V}^{t} \sum_{j=1}^{J} U_{jvtp} \geq Q_{tp}, \quad \text{all } t, p . \tag{17}$$

Note that the "guarantee conditions" have to be handled through a separate calculation, since X_{jv} is not an endogenous decision variable in the simulation model. In addition, there are the hydro-energy constraints. Let $j = h$ denote hydro, and let the decision variable S_{hvt} be the energy stored at the beginning of period t in the hydro station of vintage v. If $\hat{S}_{hvt}$ is the storage capacity of the reservoir, $0 \leq S_{hvt} \leq \hat{S}_{hvt}$. Also let W_{hvt} be the water inflow in period t, expressed in energy units and adjusted for losses due to seepage and evaporation. Then the water stored at the end of the period plus the water used for generation during the period must be less than or equal to the water stored at the beginning plus the inflow:

---

[49] See [52].

$$S_{hvm,t+1} + \sum_{p=1}^{P} U_{hvtp} \cdot \theta_p \leq S_{hvt} + W_{hvt}, \quad \text{all } h, v, t\,, \qquad (18)$$

and these constraints will be satisfied with equality if there is no spillage.

The above linear program can be solved using standard computer programs. It can be extended in various ways, as will be shown later in this paper, to include multi-purpose schemes, regional decision variables, and transmission losses. The objective function is also separable, so that nonlinearities in the cost coefficients can be treated by separable programming.

A difficulty with LP simulation models is the large number of constraints which must be satisfied in any realistic formulation of a problem. Constraints (16) in particular can become exceedingly numerous if the load duration curve is broken down into many periods and the types and vintages of plant are many. But it is possible to overcome this difficulty. Since the capital structure is predefined and fixed, these constraints form an upper bound set and can be treated by bounded variable LP methods.[50] For predominantly thermal systems the problem can be decomposed into several independent and much smaller linear programs (e.g., one for each year) since the operating decisions for one year are to a good approximation independent of those of previous years. Although LP and non-LP simulation models for planning calculations have not been reported as frequently as the other models we have discussed, it is interesting to note that they are used extensively by engineers for "real-time" load dispatching calculations.[51]

5. Global models

■ One difficulty with marginal analysis is the large number of marginal changes to a basic plan that must be considered. If it were only necessary to consider investment decisions to be made at the present time, the number of marginal changes might not be too many (although this is not true if regional variables and transmission are included in the model). However, this is not the situation; investment decisions over time must be considered, and this adds enormously to the dimensions of the problem. It is the function of global models to overcome this second difficulty with marginal analysis. Specifically, they are designed to scan and cost a large number of present and future investment policies and select the optimum. For each investment policy they cost, they must of course simulate system operation and calculate optimum operating schedules and costs. Simulation models are therefore a special case of global models.

---

[50] See Hadley [37].

[51] Articles are frequently published on this topic by the Institute of Electronics and Electrical Engineers (IEEE) (U. S.) and the Institute of Electrical Engineers (U. K.). See also Cory and Sasson [19], Ariatti, Grohmann, and Venturini [3], Vol. 1, Farmer, James, and Wells [26], and Tyren [81], Vol. 2. The well-known study of Kirchmayer [46] used Lagrange multipliers to solve the load dispatching problem.

This does not mean however that simulation models are superseded by global models. To arrive at a uniquely optimum policy in one computer run is perhaps asking too much. A formulation of a global model necessarily entails approximation. But once approximate global solutions have been reached, they can be, and are, examined in more detail by a marginal analysis using simulation models. In this way, as Bessière and Petcu have argued, global models and marginal analysis using simulation models are complementary techniques.[52]

The first global models to be developed were formulated as linear programs. In recent years, some workers have turned to non-LP formulations on the grounds that they are computationally more efficient. We shall consider (1) linear programming, (2) non-linear programming, and then return to (3) an LP reformulation of the global models which appears to be at least as computationally efficient as the non-LP formulations.

□ **(1) Linear programming.** The following formulation borrows in particular from Massé, Bessière, and Whitting and Forster.[53] It is very similar to the formulation in Section 2, except that, as with the LP simulation model, we break down the load duration curve into $p = 1 \cdots P$ discrete blocks. Adding the capital cost terms to the generation costs, the investor's objective is to choose the investments X_{jv} over $v = 1 \cdots T$ and the associated operating decisions U_{jvtp} over $t = 1 \cdots T$ and $p = 1 \cdots P$, so as to minimize total discounted system costs.

$$\text{Minimize} \sum_{j=1}^{J} \sum_{v=1}^{T} C_{jv} X_{jv} + \sum_{j=1}^{J} \sum_{t=1}^{T} \sum_{v=-V}^{t} \sum_{p=1}^{P} F_{jtvp} \cdot U_{jtvp} \cdot \theta_p \qquad (19)$$

subject to the following constraints:

(1) The plant in operation must be sufficient at all times to meet the instantaneous power demand:

$$\sum_{j=1}^{J} \sum_{v=-V}^{t} U_{jtvp} \geq Q_{tp}, \qquad \begin{aligned} t &= 1 \cdots T \\ p &= 1 \cdots P. \end{aligned} \qquad (20)$$

(2) The output of each plant must not be greater than the available capacity. In general, the available capacity is somewhat lower than actual capacity on account of planned outage (maintenance) and unplanned outage (faults). If the availability factor for plant X_{jv} in year t is a_{jvt}, this constraint is then:

$$U_{jtvp} \leq a_{jvt} \cdot X_{jv}, \qquad \begin{aligned} j &= 1 \cdots J \\ v &= -V \cdots t \\ t &= 1 \cdots T \\ p &= 1 \cdots P. \end{aligned} \qquad (21)$$

---

[52] See [11].

[53] See [62], [10], and [29,30] respectively.

(3) There will be an upper limit to the hydro energy available in any period t. Let this limit for each vintage be H_{vt}. Then:

$$\sum_{p=1}^{P} U_{htvp}\cdot\theta_p \le H_{vt}, \qquad \begin{aligned} t &= 1\cdots T \\ v &= V\cdots t \\ j &= h\,(\text{hydro})\,. \end{aligned} \tag{22}$$

(4) Equality constraints represent the initial capital stock. Let the plant initially on the system be denoted by $\hat{X}_{jv}$, $j = 1\cdots J$ and $v = -V$ to 0. Then:

$$X_{jv} = \hat{X}_{jv}, \qquad \begin{aligned} j &= 1\cdots J \\ v &= -V\cdots 0. \end{aligned} \tag{23}$$

(5) To guarantee peak power supplies to an acceptable probability limit, the installed capacity must be sufficient to meet the mean expected demand with a margin of reserve capacity (m) to allow for demands above the mean expectations:

$$\sum_{j=1}^{J}\sum_{v=-V}^{t} X_{jv} \ge Q_{tp}\cdot(1+m), \qquad \begin{aligned} t &= 1\cdots T \\ p &= 1\,. \end{aligned} \tag{24}$$

Similarly, there may be a guarantee condition for energy supplies, requiring a constraint similar to (7). A proper study of this condition will generally require a seasonal model, which adds substantially (but not prohibitively) to the dimensions of the model. Note however that the effect of (7) is essentially to limit the ratio of hydro to thermal plant on the system. As a shortcut therefore, but one evidently involving assumptions which may sometimes be rather approximate, an annual model can still be used but with a restriction on the ratio of hydro to thermal plant on the system. This ratio can be ascertained from separate reserve and reliability studies.

(6) Finally, there are a number of "local" and other constraints. For example, the number of hydro sites may be limited, certain decisions may be political, the future investment program may have been partly determined by previous studies, and so on. Constraints to represent shortages of capital and foreign exchange may also be introduced. All decision variables are of course non-negative.

The constraint matrix has a very simple form. The coefficients are mainly zeros and ones, and fall into regular patterns. Matrix generator programs can be written to produce the constraint matrices. This cuts down on data and input preparation considerably. Since the constraint matrix is not very dense, computation can be very fast. Although this is the simplest form of the global model, much has been, and can be, accomplished with it, and the use of linear programming and its extensions by the electricity supply industry is common practice in many countries. Bessière[54] informs us that LP formulations were first studied for Electricité de France in the mid-1950s by Massé and Gibrat, who published their results in 1957.[55]

---

[54] In [9].

[55] In [63]. This was soon followed by a paper by Massé and Bessière [10], a comparison of which with the following papers will show that the principles

The above model can be extended in many directions to include, for example:

(1) Optimum replacement;[56]

(2) Optimum locations of plant and directions of bulk energy transmission;[57]

(3) Optimum storage capacities and storage policies for hydroelectric plant, including special constraints on the operation of multi-purpose hydro schemes involving electricity generation;

(4) Integer variables to represent the large fixed-cost component of hydro and nuclear schemes and transmission equipment;[58] and

(5) Nuclear fuel cycling.[59]

Examples of the first three extensions are presented in Section 6. The simulation models corresponding to each of these extensions follow in a rather obvious way by pre-defining all the capacity variables and using the LP to search for optimum dispatching schedules only.

A difficulty with LP models of the above type is the large number of constraints encountered in any realistic formulation of a problem. The principal cause of this is constraint (21): we must ensure that the output of every plant on the system in every year of the study and on every interval of the load duration curve does not exceed its maximum available capacity. This is the same problem as was raised in connection with the LP simulation models; this time however, the *X*'s are not constants but decision variables, and the constraint cannot therefore be handled by bounded variable LP methods. If there are on the average 20 plants on the system, and we break the load duration curve into ten discrete intervals and take a 30-year period broken down into 6 by 5-year intervals, then we have about 1,200 constraints of this type and about 1,500 constraints in the problem. This is quite a large linear program, although standard computer programs are now available which can handle up to 10,000 constraints with mixed integer/continuous variable facilities;[60] and the use of matrix generators and "report writers" make data preparation and output processing quick and simple.

☐ **(2) Non-linear programming.** In the early 1960s computers could not handle anything like this number of constraints. Bessière and

---

remain the same, whatever the country and whatever the date: Whitting and Forster [29,30]; Nitu et al. [67]; United Nations Symposia [82,84]; and papers by Eibenschuz [24], Frankowski [31], and Deonigi [20] in the 1970 Symposium held by the International Atomic Energy Authority (IAEA) in Vienna. Among the richest sources of information known to the present writer are the Power Systems Computation Conferences (PSCC) held in London (1963), Stockholm (1966), and Rome (1969), published by the Department of Electrical and Electronic Engineering, Queen Mary College, University of London.

[56] See Massé [62].

[57] See Fernandez, Manne, and Valencia [28].

[58] See Gately [33] and Fernandez, Manne, and Valencia [28].

[59] See Frankowski [31].

[60] This is the capacity of the OPHELIE II LP System for the Control Data Corporation 6600 Computer. We have made considerable use of this very powerful system at the IBRD via a remote batch terminal connected to the CDC CYBERNET nationwide computer system.

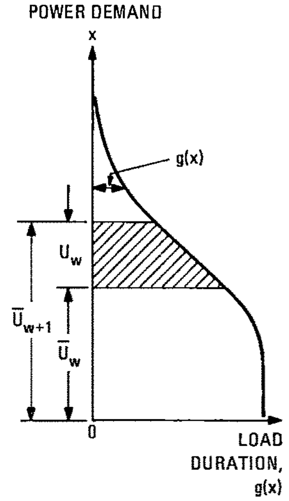
Albert and Larivaille[61] report that in 1958, 180 constraints and 200 unknowns approached the maximum that computers could handle at that time; apparently the main reason why Electricité de France turned to non-linear programming was to overcome this constraint problem. For the all-thermal system of the U. K., Phillips et al. also developed a non-LP model;[62] it will be described here briefly in order to show how non-linear programming overcomes to a large extent the constraint problem.

The idea is to prearrange all the plants that are or may be connected to the system in any year in "merit order" in the data input. That is, the operating sequence is decided in advance by inspecting the marginal operating costs before the computer run is commenced. By this device, all the operating variables and their associated capacity constraints can be satisfied implicitly and deleted from the formulation.

To reduce notation, we shall drop the subscript t until needed. Moreover, we shall represent plant type j, vintage v, by a single subscript, w, where $w = 1, 2, 3 \ldots W$, and W is the total number of plant of all vintages in year t. We let X_w be the *available* power capacity of thermal plant w, U_w its power output at any instant, and F_w its operating cost. The ordinate on the demand duration curve (Figure 6) is denoted by x, where x is of course in units of power demand, and we denote the duration of demand x by $g(x)$. Now define the subscripts w such that their sequence locates the plant in merit order, as follows:

FIGURE 6

LOAD DURATION CURVE WITH NOTATION FOR A NON-LINEAR PROGRAMMING MODEL



U_w = OUTPUT OF PLANT w

$\bar{U}_w = \sum_{w'=1}^{w} U_{w'}$

$$0 \leq F_1 \leq F_2 \leq F_3 \cdots \leq F_w \cdots \leq F_W, \tag{25}$$

where F_w is the new notation for the operating costs of plant j, vintage v, in year t.

The cost of operating plant w in merit order is then given by

$$\int_{\bar{U}_{w-1}}^{\bar{U}_w} F_w \cdot g(x)dx = F_w(G(\bar{U}_w) - G(\bar{U}_{w-1})), \tag{26}$$

where

$$G(\bar{U}_w) = \int_0^{\bar{U}_w} g(x)dx, \quad \text{and} \quad \bar{U}_w = \sum_{w'=1}^{w} U_{w'} .$$

Adding (26) over $w = 1$ to W gives the total operating cost (TOC) to be[63]

$$(\text{TOC}) = \sum_{w=1}^{W} (F_w - F_{w+1})G(\bar{U}_w),$$

which, after substitution for $\bar{U}_w$, becomes

$$(\text{TOC}) = \sum_{w=1}^{W} (F_w - F_{w+1})G(\sum_{w'=1}^{w} \bar{U}_{w'}) . \tag{27}$$

A reasonable simplifying assumption can be made to further reduce the size of the problem:[64] that the available plant capacity

---

[61] In [9] and [1], respectively.

[62] See [69].

[63] Note that $G(\bar{U}_0) = 0$, and we use the convention $F_{W+1} = 0$.

[64] This assumption is not made by Phillips et al. [69].

X_w will be operated at full power in the interval w to $w + 1$, so that $U_w = X_w$ exactly. Therefore,

$$(\mathrm{TOC}) = \sum_{w=1}^{W} (F_w - F_{w+1})G(\sum_{w'=1}^{w} X_{w'}) \,. \tag{28}$$

Reintroducing the subscripts j, v, and t into the formulation, we find that the investor's objective is

$$\text{Minimize } \sum_{j=1}^{J} \sum_{v=1}^{T} C_{jv}\cdot X_{jv} + \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{v=-V}^{t} (F_{j,v,t} - F_{j,v+1,t})G_{jvt} \,, \tag{29}$$

where

$$G_{jvt} = G_t(\sum_{i=1}^{j} \sum_{u=-V}^{v} X_{iu})$$

$$\text{Subject to } \sum_{j=1}^{J} \sum_{v=-V}^{t} X_{jv} \geq \hat{Q}_t, \qquad t = 1, \ldots T \,, \tag{30}$$

where $\hat{Q}_t$ is the peak power demand in year t.

Thus it is possible to represent the operating cost explicitly in terms of the plant capacities of the system, their operating costs, and the shape of the load duration curve. No special algorithm is needed to schedule the plant optimally. All variables U_{jtvp} satisfy the operating capacity constraints implicitly. Moreover, the demand constraints are satisfied implicitly except at times of peak demand. The formulation thus accomplishes an enormous reduction in the number of constraints to be satisfied—at the expense, however, of a complex, nonseparable, but convex objective function.

Apparently this model is now in constant use in the U. K. for the evaluation of investment plans. The non-linear program of Electricité de France has also been in use for several years.[66]

The problem with turning to non-linear programming is that we lose very considerable advantages of LP computer software, which include flexibility, very versatile management processing systems, input and output processors, and integer facilities. Also not to be underestimated is the fact that the LP formulations are simpler and can be readily rewritten to cover other problems such as replacement, bulk electrical energy transmission, hydro storage policy, and multi-purpose projects.

The question arises then, Can we retain an LP form and yet reduce the constraint problem? The answer is that we can. The reason why the non-LP model reduces the number of constraints is not because it is intrinsically more efficient than linear programming; it is because in the non-LP model we include *a priori* information about system operating characteristics which we exclude from LP models. It is because this information is excluded that we get so many constraints in the LP model. By including it, we can reduce the problem size to virtually the same numerical proportions we encounter with non-LP forms.

---

[66] See Bessière [8], Albert and Larivaille [1], and the references in Bessière's review [9].

☐ **(3) Recovering an LP form.** The way we include the information is as follows.[66] We know that as we move along the time axis of the load duration curve the output of any plant will not be increased. In fact it will either remain the same as it was before, or it will be reduced. Suppose that we define new operating decision variables, Z's, to replace the U's, which represented the output of each plant. These Z's (which we will call Z-substitutes) are defined to be the *decrease* in output of any plant as we move along the load duration curve. For any plant j, v:

$$Z_{jtvp} = U_{jtvp} - U_{jtv,p+1} \geq 0, \qquad p = 1 \cdots P - 1 \qquad (31)$$

with

$$Z_{jtvP} = U_{jtvP} \geq 0. \qquad (32)$$

As we move along the load duration curve (that is, as load decreases) the power output of plant j, v, is never increased;[67] it follows that the sum of power reductions from $p = 1$ to P is less than the available power capacity of plant j, v. Hence

$$\sum_{p=1}^{P} Z_{jtvp} \leq a_{jvt} \cdot X_{jv}, \qquad \text{all } j,\ t,\ v\,. \qquad (33)$$

This constraint, together with the non-negativity constraints on Z_{jtvp}, are necessary and sufficient conditions for constraints (21) to be satisfied. First, in view of (32), U_{jtvP} is non-negative if Z_{jtvP} is non-negative, and it follows from (31) that if Z_{jtvp} is non-negative for all p, so must be U_{jtvp}. Second, in view of (33) no combination of the values of Z_{jtvp} can exceed $a_{jvt} \cdot X_{jv}$. From (31) and (32) we derive the relation that

$$U_{jtvp} = \sum_{p'=p}^{P} Z_{jtvp'} \leq a_{jvt} \cdot X_{jv}\,. \qquad (34)$$

Hence U_{jtvp} cannot exceed $a_{jvt} \cdot X_{jv}$ if (31), (32), and (33) are satisfied.

We can thus replace the U's, which had to satisfy plant capacity constraints on every portion of the load duration curve, with new non-negative variables, the Z-substitutes, which satisfy one constraint for the whole curve. Approximately, the number of constraints is reduced by $1/P$. The 1,500-constraint problem we mentioned earlier[68] is now reduced to 150 constraints. It has also been our experience that, although the density of the constraint matrix is increased, computing times have been reduced by a factor of 2 or more (sometimes by a factor of 5).

This is a useful result for those who prefer to work with the simpler LP models. For those directly involved with investment planning in the industry, it means that computational problems are kept to a manageable size, without loss of generality and with many advantages in terms of computer software. If the cost coefficients are non-linear (as happens for example with studies of hydro resources), the objective function is separable so that non-linearities can be treated by interpolation and the LP form recovered. For those involved with economic research it means that global models based on linear programming can provide a realistic view of the investment problem—at least from the supply side. If demand is introduced as an endogenous variable, however, we have to return to non-linear programming, particularly if peak and off-peak demands are considered interdependent since the objective function is then no longer separable.

---

[66] I am grateful to Ivan Whitting of the National Gas Council, U. K., for pointing out the following device to me. The idea was apparently suggested by E. M. L. Beale during a conversation. It has apparently not been published previously.

[67] Unless seasonal variations in hydro flows and maintenance schedules are important. In such a case it is necessary to use one load duration curve for each season, and the above statement holds once again for each load duration curve.

[68] Section 5(1)

6. Three extensions to LP models

■ This contraction in the size of the problem, together with the rapid expansion in the size of linear programs that can be handled on computers, also enables us to expand the detail and content of both global and simulation models. Below we look at three extensions that are being worked on at the International Bank for Reconstruction and Development: replacement, an approximate treatment of transmission, and investment and operating decisions in systems with hydro storage schemes.

☐ **(1) Inclusion of replacement.** Optimum replacement of a power station usually occurs when it is cheaper to expand and operate the power system without this power station. This may arise because of rising operating and maintenance costs relative to those of new plant, or because sites for new power stations are short and old ones need to be scrapped to make room for new and larger ones.

An accurate treatment of replacement requires explicit separation of fixed annual operation and maintenance costs from other costs. Usually these costs are added to the annuitized charges on capital, while the variable maintenance costs are added to the other variable operating costs. Following Massé[69] we could define a new decision variable X_{jv}^* to represent the plant scrapped of type j, vintage v, and write the capital cost terms in the objective function as $C_{jv}(X_{jv} - X_{jv}^*)$, where C_{jv} equals the fixed annual costs plus annuitized charges on capital. This can however lead to solutions which scrap plant before capital costs have been accounted for by the annuities. Since the decision to scrap new plant requires a new decision variable, we can associate this variable directly with fixed maintenance and operating costs.

We denote by M_{jvt} the discounted, fixed maintenance, and operating costs of plant type j, vintage v, in year t. The problem is to decide how much of this plant should remain in service in year t. Let R_{jvt} be the amount of type j, vintage v, remaining in service in year t. The objective and the constraints follow much the same pattern as before. Using the same notation,

Objective

$$\text{Minimize}\begin{cases} \sum_{j=1}^{J} \sum_{v=1}^{T} C_{jv} \cdot X_{jv} + \sum_{j=1}^{J} \sum_{t=1}^{T} \sum_{v=-V}^{t} M_{jtv} \cdot R_{jtv} \\ \quad + \sum_{j=1}^{J} \sum_{t=1}^{T} \sum_{v=-V}^{t} \sum_{p=1}^{P} F_{jtvp} \cdot U_{jtvp} \cdot \theta_p \, . \end{cases} \tag{35}$$

---

[69] [62], p. 187 *et seq.*

Constraints

Total plant remaining greater than peak demand requirements (plus an allowance for reserves),

$$\sum_{j=1}^{J}\sum_{v=-V}^{t} R_{jtv} \geq Q_{tp}\cdot(1+m) \qquad t = 1\cdots T, \quad p = 1 \tag{36}$$

Plant types j, v, remaining less than plant installed of type j, v,

$$R_{jtv} \leq X_{jv} \qquad j = 1\cdots J, \quad t = 1\cdots T, \quad v = 1\cdots t, \tag{37}$$

Plant remaining of given type j, v, never increases,

$$R_{j,t+1,v} \leq R_{jtv} \qquad j = 1\cdots J, \quad t = 1\cdots T, \quad v = 1\cdots t, \tag{38}$$

Total plant operating always sufficient to meet demand,

$$\sum_{j=1}^{J}\sum_{v=-V}^{t} U_{jtvp} \geq Q_{tp} \qquad t = 1\cdots T, \quad p = 1\cdots P, \tag{39}$$

A plant's output never exceeds remaining available capacity,

$$U_{jtvp} \leq a_{jtv}R_{jtv} \qquad j = 1\cdots J, \quad t = 1\cdots T, \quad v = 1\cdots t, \quad p = 1\cdots P, \tag{40}$$

Restrictions on energy available from each hydro plant,

$$\sum_{p=1}^{P} U_{htvp}\cdot\theta_p \leq H_{vt} \qquad t = 1\cdots T, \quad v = -V\cdots t. \tag{41}$$

Finally, there are "local" and non-negativity constraints, and initial conditions. Z-substitutes can of course be used in this formulation. Constraints (36) are the "guarantee conditions" for peak power; constraints to guarantee energy supplies can also be introduced in the ways previously discussed.

☐ **(2) Approximate inclusion of transmission.**[70] Transmission systems reduce costs of supply in four ways. First, if regions are interconnected, in the event of generator failure in one region, it is possible to call upon the reserves of other regions; aggregate reserve capacity with interconnection is less than is required without interconnection. Second, if peak demands occur at different times in different regions, then interconnection permits peak power capacity to be exported and imported, and the aggregate peak demand can be met with less capacity. Third, if the transmission system is designed to transmit energy in large quantities, the markets of regions rich in energy resources (fossil or hydro) can be expanded, and regions less rich can import the cheaper energy. Fourth, with interconnection, larger units can be installed embodying considerable economies of scale. These four aspects of the transmission system—the pooling of reserve capacity, the pooling of peak capacity, the opening of markets to regions rich in low-cost resources, and economies of scale—can make large savings as compared to the costs of the transmission lines and transmission losses.

---

[70] I am very grateful to Mr. Narong Thananart, who corrected some mistakes in the original formulation and who wrote a matrix generator for use in case studies.

In the following formulas, regions of generation are denoted by integers $g = 1 \cdots G$, and regions of demand by integers $d = 1 \cdots D$. Y_{gdv} denotes the increment of transmission capacity (expressed in MW) connecting g, d, installed in year v; L_{gdv} is the discounted cost per MW installed of this increment. The power delivered to region d by station type j, vintage v, in period t, p, from region g is denoted by U_{jtvpgd}; and its total output is $\sum_d U_{jtvpgd}$. The capital and operating costs are as before, except that they must now be summed over all regions. A cost term is also required for the transmission of capital costs.[71] The objective function is therefore

$$\text{Minimize}\begin{cases} \displaystyle\sum_{g=1}^{G} \sum_{j=1}^{J} \sum_{v=1}^{T} C_{jvg} \cdot X_{jvg} \\ \displaystyle\quad + \sum_{g=1}^{G} \sum_{d=1}^{D} \sum_{v=1}^{T} L_{gdv} \cdot Y_{gdv} \\ \displaystyle\quad + \sum_{g=1}^{G} \sum_{d=1}^{D} \sum_{j=1}^{J} \sum_{t=1}^{T} \sum_{v=-V}^{t} \sum_{p=1}^{P} F_{jtvpgd} \cdot U_{jtvpgd} \cdot \theta_p \,. \end{cases} \tag{42}$$

The first constraint to be satisfied is the peak power guarantee condition that the installed capacity must be greater than the peak load presented at the power stations' terminals by a margin m, to allow for demands above mean expectations.[72] We also introduce a "diversity factor" c, which is the ratio of the aggregate peak demand to the arithmetic sum of the regional peak demands. The capacity requirements are then

$$\sum_{g=1}^{G} \sum_{j=1}^{J} \sum_{v=-V}^{t} X_{jvg} \geq (1 + m)c \sum_{g=1}^{G} \sum_{d=1}^{D} \sum_{j=1}^{J} \sum_{v=-V}^{t} U_{jtvpgd} \,, \qquad \begin{aligned} t &= 1 \cdots T \,. \\ p &= 1 \,. \end{aligned} \tag{43}$$

The diversity factor might have to be modified if the solutions suggest a different pattern of connections than the pattern used for computing the factor. Transmission capacity must also be sufficient to carry the peak load transfers:

$$\sum_{v=-V}^{t} Y_{gdv} \geq (1 + m) \sum_{j=1}^{J} \sum_{v=-V}^{t} U_{jtvpgd} \,, \qquad \begin{aligned} t &= 1 \cdots T \\ p &= 1 \\ g &= 1 \cdots G \\ d &= 1 \cdots D \,. \end{aligned} \tag{44}$$

Next, the plants' output must meet the demand and the transmission losses. If b_{gd} is the per-unit power attenuation between g, d, then the plant must be operated in each period p such that[73]

$$\sum_{v=-V}^{t} \sum_{g=1}^{G} \sum_{j=1}^{J} U_{jtvpgd}(1 - b_{gd}) \geq Q_{dtp} \,, \qquad \begin{aligned} t &= 1 \cdots T \\ p &= 1 \cdots P \\ d &= 1 \cdots D \,. \end{aligned} \tag{45}$$

---

[71] We do not treat replacement in this model, although it is obviously quite possible to do so if required. Integer variables can also be introduced to represent the indivisibilities of transmission investments, but this is not done below.

[72] Of course, there may also be a guarantee condition for energy supplies, which we do not list here.

[73] The power attenuation will vary with the square of the load transfer between g, d. This would make constraint (45) quadratic. We can only retain linearity by taking b_{gd} as a weighted average value (in fact a "mean-square" average).

Next we have the constraints that no plant can be operated above its peak available capacity:

$$c \cdot \sum_{d=1}^{D} U_{jtvpgd} \leq a_{jtvg} \cdot X_{jvg} , \quad \begin{aligned} j &= 1 \cdots J \\ v &= -V \cdots t \\ t &= 1 \cdots T \\ p &= 1 \cdots P \\ g &= 1 \cdots G . \end{aligned} \tag{46}$$

(Again, these constraints can be reduced by $1/P$ by the use of non-negative Z-substitutes.) The hydro-energy constraint takes the same form as before:

$$c \cdot \sum_{d=1}^{D} \sum_{p=1}^{P} U_{htvpgd} \cdot \theta_p \leq H_{vtg} , \quad \begin{aligned} t &= 1 \cdots T \\ v &= -V \cdots t \\ g &= 1 \cdots G . \end{aligned} \tag{47}$$

Finally, there are "local," budget, foreign exchange, and nonnegativity constraints; constraints to represent the initial conditions; and a constraint to guarantee energy supplies.

☐ **(3) Inclusion of water-storage capacity and operating policy variables.**[74] We now formulate the model to search for least-cost, evolving investment programs to satisfy an exogenous demand for electricity, a planned delivery of water to irrigation, and a planned degree of flood control. The peak storage capacity and the amount stored in each season are treated as decision variables. The model can easily be couched in a regional context, as has been done elsewhere,[75] to allow for the strong regional dependence of hydro resources and the high economies of scale which they may yield if the interconnected system is large enough to absorb them. However, to reduce notation, this is not done below; the principles are in any case identical to the ones presented in subsection (2) above.

Thermal schemes will now be denoted by subscript $j = 1 \cdots J$ and hydro storage schemes by $h = 1 \cdots H$. As before, maximum power capacities will be denoted by the decision variable X, the instantaneous outputs by U, and incremental capital and generation costs by C and F, respectively. Each year will be denoted by t and divided into $m = 1 \cdots M$ periods, which can represent months, seasons, weeks, etc., according to the accuracy desired. The demands within each period m will be represented by a load duration curve divided into $p = 1 \cdots P$ blocks. We have the following additions to notation for the hydro schemes:

$\hat{S}_{hv}$ = decision variable representing the maximum storage capacity (expressed in energy units) of scheme h, v;

K_{hv} = corresponding incremental capital cost of providing the storage capacity;

S_{htvm} = decision variable representing the actual water in storage (expressed in energy units) of scheme h, v, at the beginning of m in year t; and

---

[74] I have benefited very much from discussions with Dr. Ralph Turvey on this model (as I also have in this paper from his writings). In particular he pointed out to me the economic significance of the Kuhn-Tucker conditions of the model.

[75] See Anderson [2].

W_{htvm} = water inflows to scheme h, v, during m of year t, expressed in energy units and corrected for losses due to evaporation and seepage.

The objective function is then:

$$\text{Minimize} \left\{ \begin{array}{l} \sum_{j=1}^{J} \sum_{v=1}^{T} C_{jv} \cdot X_{jv} + \sum_{h=1}^{H} \sum_{v=1}^{T} C_{hv} \cdot X_{hv} \\ + \sum_{h=1}^{H} \sum_{v=1}^{T} K_{hv} \cdot \hat{S}_{hv} \\ + \sum_{j=1}^{J} \sum_{t=1}^{T} \sum_{v=-V}^{t} \sum_{m=1}^{M} \sum_{p=1}^{P} F_{jtvmp} \cdot U_{jtvmp} \cdot \theta_p \\ + \sum_{h=1}^{H} \sum_{t=1}^{T} \sum_{v=-V}^{t} \sum_{m=1}^{M} \sum_{p=1}^{P} F_{htvmp} \cdot U_{htvmp} \cdot \theta_p \, . \end{array} \right. \tag{48}$$

Note that the last term is small for hydro schemes, and also that there is no cost term associated with actual water in storage, S_{htvm}.

The following constraints are exactly analogous to the ones presented above: (1) installed capacity must be greater than or equal to the annual peak demand plus a margin for reserves; (2) there must be sufficient energy reserves to meet energy demand in dry seasons; (3) the aggregate plant output must meet the instantaneous power demand at all times; (4) no plant can be operated above its peak available capacity [again, (4) can be reduced by $1/P$ using Z-substitutes], and (5) "local" constraints, initial conditions, etc.

The additional constraints introduced by storage hydro are as follows. First, the water stored at the end of period m (beginning of $m + 1$) plus the water used for generation is less than or equal to the initial storage plus the inflow:

$$S_{htv,m+1} + \sum_{p=1}^{P} U_{htvmp} \cdot \theta_p \leq S_{htvm} + W_{htvm} \, , \tag{49}$$

for $m = 1 \cdots M - 1$; and for $m = M$:

$$S_{h,t+1,v,1} + \sum_{p=1}^{P} U_{htvMp} \cdot \theta_p \leq S_{htvM} + W_{htvM} \, , \tag{50}$$

where (49) and (50) must of course be satisfied for all h, t, v, and m.

If a hydro scheme is multi-purpose, involving irrigation and flood control, there will be further restrictions on the timing and the rate of energy output. Suppose the water requirements of irrigation in period m of year t are I_{htvm}. There will still be considerable flexibility in the pattern of discharge within the period m (e.g., choice between night and day discharges or between weekdays and weekends); but the aggregate amount of water discharged through turbines must at least be equal to the requirements of irrigation:[76]

$$\sum_{p=1}^{P} U_{htvmp} \cdot \theta_p \geq I_{htvm}, \quad \text{all } h, t, v, m \, . \tag{51}$$

---

[76] We also assume in this example that all the water, other than that lost by spillage and seepage, is discharged through turbines. If desired the assumption can be relaxed.

Flood control, on the other hand, sets an upper limit to the rate of discharge in certain periods. Let D_{htvm} represent this limit in period m. Then the water discharged from the hydro, minus the quantity diverted to irrigation, must not exceed this limit:

$$\sum_{p=1}^{P} U_{htvmp}\cdot\theta_p \leq I_{htvm} + D_{htvm}, \quad \text{all } h,\ t,\ v,\ m\,. \tag{52}$$

This completes the present formulation. The approach is very flexible and new features can readily be introduced. If hydro schemes are large, they can be expanded in stages instead of being introduced in one period, v. Variable head schemes, pumped storage schemes, and multi-dam cascade schemes can also be given a full analysis. Approximations can be introduced in the representation of the costs of electric power from thermal stations so as to allow room for more detail elsewhere—for example, in the representation of multi-dam schemes.[77]

Appendix

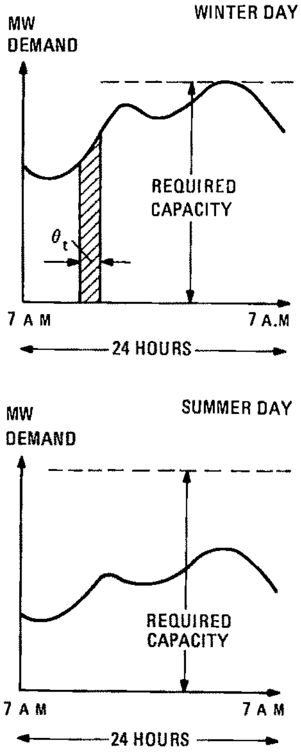
A note on the load duration curve[78]

■ The difficulty of calculating optimum operating schedules and costs is presented by the high variability of power demand, which varies throughout the day and throughout the year (see Figure 7). The operating costs are the area under this curve weighted at each time interval θ_t by the fuel costs and the outputs of the plant in that interval. To simplify the calculation of operating costs it is usual to construct a curve known as the load duration curve. This curve is constructed from the above demand curve by rearranging each load for each time interval θ_t to occur in descending order of magnitude (see Figure 8). The operating costs are thus the area under the load duration curve again weighted at each time interval by the operating costs per unit energy output and the output of each plant operating in that interval. The load duration curve makes integration of costs less difficult because it can be represented by simpler functions than the curves of Figure 7. It is a convenient form to check approximations to the patterns of demand shown in Figure 7; and we can also use Z-substitutions if we use the load duration curve.

Use of the load duration curve for calculations of operating costs introduces one important assumption: *that the costs and availability of supply depend only upon the magnitude of the load and not on the time at which the load occurs.* This assumption is quite accurate for all-thermal systems (although plant availability, because of maintenenace schedules, is seasonal), but it is only approximate for hydro schemes. To analyze hydro operation accurately it is often necessary to construct a separate load duration curve for each season and sometimes each month; if this is done the assumptions of the load duration curve are tenable again.

FIGURE 7

VARIABILITY OF POWER DEMAND



References

1. Albert, M. and Larivaille, Mme. P. "Utilisation de Modeles Globaux pour le Choix des Operations des Programmes E. D. F." Conf. I.F.O.R.S.–H.E.L.O.R.S. La Recherche Operationnelle et l'Energie Electrique, Athens, November 1968.

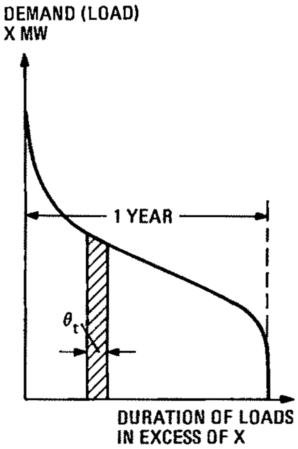
---

[77] See Hufschmidt and Fiering [42] and Maass et al. [55].

[78] See also Berrie [6], Openshaw-Taylor [68], and Kirchmayer [46].

FIGURE 8

LOAD DURATION CURVE (FOR ONE YEAR'S DEMAND)



2. Anderson, D. "Investment Analysis in Electricity Supply Using Computer Models." Working Paper No. 91, Economics Department, World Bank, 1970.
3. Ariatti, P., Grohmann, D., and Venturini, D. "A Method for Economic Load Dispatching in a Thermal Power System." Third PSCC, Rome, 1969.
4. Bernard, P. J., Dopazo, J. F., and Stagg, G. W. "A Method for Economic Scheduling of a Combined Pumped Hydro and Steam Generating System." IEEE Transactions on Power Apparatus and Systems, Vol. 83, 1964.
5. Berrie, T. W. "Further Experience with Simulation Models in System Planning." Second PSCC, Stockholm, June 1966.
6. ———. "The Economics of System Planning in Bulk Electricity Supply." *Electrical Review*, Vol. 181 (September 15, 22, 29, 1967). Reprinted in R. Turvey, ed., *Public Enterprise* [78].
7. ———, and Whitting, I. J. "The Exploration of Alternative Plans for an Expanding Electrical Power System." First PSCC on Digital Computation for Electric Power Systems, Queen Mary College, London, 1963.
8. Bessière, F. "Le modele 'Investments 85' d'Electricité de France." *Revue Francaise de l'Energie*, No. 182, 1966, p. 568 *et seq.*
9. ———. "Methods of Choosing Equipment at Electricité de France: Development and Present Day Concept." *European Economic Review*, Vol. 1, No. 1 (Winter 1969), pp. 199–211.
10. ———, and Massé, P. "Long-Term Programming of Electrical Investments," in J. R. Nelson, *Marginal Cost Pricing in Practice* [65].
11. ———, and Petcu, M. "Analyse Margenales et Optimisation Structurelle des Investissements: Application au Secteur de L'Electricité." *R.I.R.O.*, No. 6, 1967, pp. 61–81.
12. ———, and Sautter, E. "Optimization and Sub-Optimization: The Method of Extended Models in the Non-Linear Case." *Management Science*, Vol. 15, No. 1 (September 1968), pp. 1–11.
13. Boiteux, M. "The Choice of Equipment for the Production of Electric Energy," in J. R. Nelson, *Marginal Cost Pricing in Practice* [65].
14. ———. "Marginal Cost Pricing," in J. R. Nelson, *Marginal Cost Pricing in Practice* [65].
15. ———. "The 'Tarif Vert' of Electricité de France," in J. R. Nelson, *Marginal Cost Pricing in Practice* [65].
16. ———, and Stabi, P. "The Determination of Costs of Expansion of an Interconnected System," in J. R. Nelson, *Marginal Cost Pricing in Practice* [65].
17. Cash, P. W. and Scott, E. C. "Security of Supply in the Planning and Operation of European Power Systems." Fourteenth International Union of Producers and Distributers of Electricity (UNIPEDE) Congress, Madrid, 1967.
18. Cazalet, E. G. "Decomposition of Complex Decision Problems with Applications to Electric Power System Planning." Stanford Research Institute, Menlo Park, California, 1970.
19. Cory, B. J. and Sasson, A. M. "Non-Linear Programming Techniques for the Load-Flow, Minimum Loss, and Economic Dispatching Problems." Third PSCC, Rome, 1969.
20. Deonigi, D. E. "A Computer Simulation for Evaluating Nuclear Power Stations in a National Electric Power Company." IAEA/SM-139/40, 1970.
21. Dhrymes, P. J. and Kurtz, M. "Technology and Scale in Electricity Generation." *Econometrica*, Vol. 32, No. 3 (July 1965), pp. 287–315.
22. Dillard, J. K. and Sells, H. K. "An Introduction to the Study of System Planning by Operational Gaming Models." IEEE Transactions, Vol. 78, Part III, 1959.
23. Eckstein, O. "A Survey of the Theory of Public Expenditure Criteria." National Bureau of Economic Research Conference on Public Finances, 1961.
24. Eibenschutz, J. "Planeacion de Centrales Nucleoelectricas." IAEA/SM-139/1, 1970.
25. Electricité de France. "L'etude a Long-Terme des Investissements a l'Aide d'un Programme Non'Lineaire: Le Modele Investissements 85." Note from Department of General Economic Studies, EDF, 1965.
26. Farmer, E. D., James, K. W., and Wells, D. W. "Computer Scheduling of Generation in a Power Supply System." Third PSCC, Rome, 1969.

27. FERNANDEZ, G. and MANNE, A. S. "A Model for Planning Investments in Generating Facilities." Internal Working Paper, Stanford University, February 1969.
28. ———, and VALENCIA, J. A. "Multi-Level Planning for Electric Power Projects." Mimeographed. Internal Working Paper, Stanford University, 1970.
29. FORSTER, C. I. K. "The Statistical Basis to National Fuel Policy." Institute of Actuaries, London, January 1969.
30. ———, and WHITTING, I. J. "An Integrated Mathematical Model of the Fuel Economy." *Statistical News No 3*, London, November 1968.
31. FRANKOWSKI, W. "Economic Integration of Nuclear Power Stations in Electric Power Systems in the Light of Optimization Research." IAEA/SM-139/10, 1970.
32. GALATIN, M. *Economies of Scale and Technical Change in Thermal Power Generation*. Amsterdam: North-Holland Publishing Co., 1968.
33. GATELY, D. "Investment Planning for the Electric Power Industry: An Integer Linear Programming Approach." Research Report 7035, Department of Economics, University of Western Ontario, London, Canada, 1970.
34. GESSFORD, J. and KARLIN, S. "Optimal Policy for Hydro-Electric Operations," in K. J. Arrow, S. Karlin, and H. Scorf, *Studies in the Mathematical Theory of Inventory and Production*, Stanford, Calif.: Stanford Univ. Press, 1958.
35. GIGUET, R. "Les Programmes d'Equipment Electrique Consideres du Point de Vue de l'Economie Appliquee." Institute de Science Economique Appliquee, No. 1, 1951.
36. GOLDSMITH, K. "The Influence of Nuclear Generation on the Planning and Operation of Interconnected Power Systems." IAEA/SM-139/15.
37. HADLEY, G. *Linear Programming*. Reading, Mass.: Addison-Wesley, 1962.
38. ———. *Non-Linear and Dynamic Programming*. Reading, Mass.: Addison-Wesley, 1964.
39. HAISSMAN, I. "Optimising the Long-Range Operation of the California Aqueduct by Incremental Dynamic Programming." Mimeographed. Operations Research Center, University of California, Berkeley, 1968.
40. HARBERGER, A. and ANDREATTA, N. "A Note on the Economic Principles of Electricity Pricing," in P. N. Rosenstein-Rodan, ed., *Pricing and Fiscal Policies*, Cambridge, Mass.: M.I.T. Press, 1964.
41. HOPKINS, D. S. P. "Sufficient Conditions for Optimality in Infinite Horizon Linear Economic Models." Technical Report No. 69-3, Stanford University, 1969.
42. HUFSCHMIDT, M. M. and FIERING, M. B. *Simulation Techniques for Design of Water Resource Systems*. Cambridge, Mass.: Harvard Univ. Press, 1966.
43. JACOBY, H. P. "Analysis of Investments in Electric Power." Mimeographed. Economic Development Series, Center for International Affairs, Harvard University, 1967.
44. JOHNSTON, J. *Statistical Cost Analysis*, New York: McGraw-Hill Book Co., 1960.
45. JONAS, P. J. "A Computer Model to Determine Economic Performance Characteristics of British Generating Stations." British Computer Conference, Brighton, 1966.
46. KIRCHMAYER, L. K. *Economic Operation of Power Systems*. New York: John Wiley & Sons, Inc., 1958.
47. KOOPMANS, T. C. "Water Storage Policy in a Simplified Hydro-Electric System." Paper No. 115, Cowles Foundation for Research in Economics, Yale University, New Haven, Connecticut, 1958.
48. KRUTILLA, J. V. and ECKSTEIN, O. *Multiple Purpose River Development*. Baltimore, Md.: Johns Hopkins Press, 1958.
49. LARIVAILLE, MME. P. "Etude des Problemes Poses pour l'insertion dans un Reseau Electriaue d'un Reacteur Nucleaire d'une Puissance Importante pour Rapport a Cell du Reseau." IAEA/SM-139/22, 1970.
50. LENCZ, I. "The Planning of the Power System Development with Mathematical Model for Two Points." Third PSCC, Rome, 1969.
51. LIEFTINCK, P., SADOVE, A. R., and CREYKE, T. C. *Water and Power Resources of West Pakistan*. 3 Vols. Baltimore, Md.: Johns Hopkins Press, 1969.
52. LINDQVIST, J. "Operation of a Hydrothermal Electric System: A Multi-stage Decision Process." IEEE Transactions, April 1962.

53. Little, J. D. C. "The Use of Storage Water in a Hydro-Electric System." Journal of the Operations Research Society of America, Baltimore, Vol. 3, No. 2 (May 1955), pp. 187–97.
54. Littlechild, S. C. "Marginal Cost Pricing with Joint Costs." *Economic Journal*, Vol. 80, No. 318 (June 1970), pp. 323–31.
55. Maass, A., Hufschmidt, M. M., Dorfman, R., Thomas, H. A., Marglin, S. A., and Fair, G. M., eds. *Design of Water-Resource Systems.* Cambridge, Mass.: Harvard Univ. Press, 1962.
56. Manne, A. S. "Linear Programming and Sequential Decisions." *Management Science*, Vol. 6, No. 3 (April 1960), p. 259–67.
57. ———. "Product-Mix Alternatives: Flood Control, Electric Power and Irrigation." Cowles Foundation Discussion Paper No. 95, October 14, 1960.
58. ———. "A Dynamic Multi-Sector, Multi-Skill Model for Mexico." Mimeographed. IBRD Research Center's Project No. 1, Memos 70-5 and 70-6, June 10, 1970.
59. Massé, P. *Les Reserves et Regulation de l'Avenir dans la Vie Economique.* 2 Vols. Paris: Hermourn, 1946.
60. ———. "Electrical Investments," in J. R. Nelson, *Marginal Cost Pricing in Practice* [65].
61. ———. "Some Economic Effects of the 'Tarif Vert'," in J. R. Nelson, *Marginal Cost Pricing in Practice* [65].
62. ———. *Optimal Investment Decisions.* Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1962.
63. ———, and Gibrat, R. "Application of Linear Programming to Investments in the Electrical Power Industry," in J. R. Nelson, *Marginal Cost Pricing in Practice* [65].
64. Morlat, G. "On Instructions for the Optimal Management of Seasonal Reservoirs," in J. R. Nelson, *Marginal Cost Pricing in Practice* [65].
65. Nelson, J. R. *Marginal Cost Pricing in Practice*, Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1964.
66. Nerlove, M. "Returns to Scale in Electricity Supply," in C. F. Christ, ed., *Measurement in Economics: Studies in Mathematical Economics and Econometrics*, Stanford, Calif.: Stanford Univ. Press, 1963.
67. Nitu, V., Bordeian, N. Goldenberg, C., and Monolescv, C. "Models for the Study of Power Systems Development by Means of Digital Computers." Third PSCC, Rome, 1969.
68. Openshaw-Taylor, E. and Boal, G. A. *Power System Economics.* London: Edward Arnold Ltd., 1969.
69. Phillips, D., Jenkin, F. P., Pritchard, J. A. T., and Rybicki, K. "A Mathematical Model for Determining Generating Plant Mix." Third PSCC, Rome, 1969.
70. Pressman, I. "A Mathematical Formulation of the Peak-Load Pricing Problem." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 304–26.
71. Prest, A. R. and Turvey, R. "Cost-Benefit Analysis: A Survey." *Economic Journal*, Vol. 75, No. 296 (December 1965), pp. 683–735. Reprinted in E. A. Robinson, *Surveys of Economic Theory*, New York: St. Martin's Press, 1966.
72. Rabar, F. "Simulation of Investment Policy in the Energy Economy." Working Paper No. 16, Research Program in Industrial Economics, Case Western Reserve University, Cleveland, Ohio, 1970.
73. Rogers, S. "A Dynamic Model for Planning Capacity Expansions: An Application to Plant Reliability and Electric Power Systems." Department of Operations Research, Stanford University, 1970.
74. Stagg, G. W. and El-Abiad, A. H. *Computer Methods in Power System Analysis.* New York: McGraw-Hill Book Co., 1968.
75. Stanford Research Institute. "Decision Analysis of Nuclear Power Plants in Electrical System Expansion." SRI Project 6469, Menlo Park, California, 1968.
76. Turvey, R. "On Investment Choices in Electricity Generation," *Oxford Economic Papers*, Vol. 15, No. 3 (1965), pp. 279–86.
77. ———. *Optimal Pricing and Investment in Electricity Supply.* London: Allen and Unwin, 1968.
78. ———. *Public Enterprise.* New York: Penguin, 1968.

79. ———. "Marginal Cost." *Economic Journal*, Vol. 79, No. 314 (June 1969), pp. 282–300.
80. ———. "Note." *Economic Journal*, Vol. 81, No. 322 (June 1971), pp. 371–5.
81. Tyren, L. "Short-Range Optimisation of a Hydro-Thermal System by a Gradient Method Combined with Linear Programming." Third PSCC, Rome, 1969.
82. United Nations. "Economic Methods and Criteria used in the Electric Power Industry." Economic Commission for Europe, ST/ECE/EP21, 1963.
83. United Nations Economic Commission for Europe. "Symposium on Problems of Multiple Purpose River Development Connected with Hydro-Electric Power Production." MAD/SYMP/EP, August 7, 1969.
84. ———. "Symposium on the Application of O.R. Methods in the Solving of Economic Problems of Planning and Operation of Large Electric Systems and on the Use of Computers for that Purpose." Varna, Bulgaria, May 25–27, 1970.
85. van der Tak, H. G. "The Economic Choice Between Hydro-Electric and Thermal Power Developments." Staff Occasional Paper No. 1, World Bank, 1966.
86. Williamson, O. E. "Peak-Load Pricing and Optimal Capacity under Indivisibility Constraints." *The American Economic Review*, Vol. 56, No. 4 (September 1966), pp. 810–27.

COMMENTS AND REVIEWS

302 **Operations research and systems analysis in the public sector**
PAUL R. KLEINDORFER

307 **The theory of peak-load pricing: a final note**
GEORGE W. WILSON

311 **Note on "gold plating" or "rate base padding"**
EDWARD E. ZAJAC

316 **Cash-flow patterns and the choice of a depreciation method**
ALAN KRAUS AND RONALD J. HUEFNER

335 **Metering costs and marginal cost pricing in public utilities**
G. FRANKLIN MATHEWSON AND G. DAVID QUIRIN

340 **A state preference approach to public utility pricing and investment under risk**
STEPHEN C. LITTLECHILD

346 **User charges as a means for pollution control: the case of sewer surcharges**
DON ETHRIDGE

354 **Errata**

Operations research and systems analysis in the public sector

Paul R. Kleindorfer

Assistant Professor of Management and Operations Research
Massachusetts Institute of Technology

■ This review considers principally the books of de Neufville and Stafford[1] and Wagner[2] with an emphasis on their contribution to the theory and practice of systems analysis in the public sector. After some general remarks on the nature of operations research and systems analysis, a rough framework is given to structure problem areas of interest. The books in question are then briefly outlined. Finally, some comments are offered with respect to their contribution in terms of the framework presented.

☐ **Operations research and systems analysis.** Traditional systems analysis has concentrated on problem definition, specification of choice alternatives, and selection methods in large-scale systems.[3] The primary applications of systems analysis began in defense programs and then expanded rapidly over the past decade to other areas in federal, state, and local government. Because of the focus of such programs it is natural that systems analysis and economics have complemented each other structurally and methodologically.

Operations research, on the other hand, has been primarily concerned with selection methods and, particularly, optimization procedures. To be sure, the focus of any good applied operations research study encompasses problem definition and alternative specification.[4] However, until recently the motivating force dominating the operations research paradigm has been problems on the operational level in private industry susceptible to exact mathematical treatment. In this context one generally finds problem formulation and alternative specification relatively easy when compared with the difficulties encountered in large-scale public sector systems.

As a result of these differences in background, a clear understanding of the relation between operations research and systems analysis has not yet developed in the minds of scientists involved in these fields. Rather than attempting to make some arbitrary partition of public sector design and evaluation problems into the respective camps, I propose to simply delineate and structure some general areas of interest emerging in the collective literature of management science in the public sector. I return to differences in focus in this field when discussing the approaches of de Neufville and Stafford and Wagner below.

---

[1] R. de Neufville and J. H. Stafford, *Systems Analysis for Engineers and Managers* (New York: McGraw-Hill Book Co., 1971).

[2] H. M. Wagner, *Principles of Operations Research* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1969).

[3] See de Neufville and Stafford, *op cit.*, Chapter 1, and references therein.

[4] For a recent discussion of the role of operations research in practice, see M. Radnor, A. H. Rubenstein, and D. A. Tansik, "Implementation in Operations Research and R-D in Government and Business Organizations," *Operations Research*, Vol. 18, No. 6 (November/December 1970), pp. 967–91.

Areas of concern in the design, planning, and control of public sector systems may be broadly classified as follows:

(1) Models of activities external to the decision-making environment: the sociological and economic environment of public sector activities.
(2) Models of activities internal to the decision-making environment: management system frameworks.
(3) Specific applications studies: coordination of external and internal models in specific fields such as health, education, and criminal justice.
(4) General methodological considerations: simulation, econometrics, and optimization in systems studies.

Under external models some important issues are collective choice and equity[5] and the measurement and evaluation of project impact on society.[6] A more recent area of interest has been general coordination issues associated with the use of citizen preference information in internal decision-making.[7]

With respect to internal models of decision-making the main thrust of research has been in the area of budgeting as a coordinating mechanism for planning and control.[8] The mixed success of PPBS[9] (Planning, Programming and Budget Systems) and recent criticism of supporting management information system approaches[10] indicate, however, that the development of a management framework for the public sector will require the more particularized function- and discipline-oriented study which has characterized research in the private sector over the past 15 years. Some of the impetus for functional area studies, e.g., in organizational studies and management information systems, will no doubt emerge from the strong current interest of schools of management in public sector decision-making. On the discipline side, the development of research interests in particular applications areas is already far advanced, as a casual glance at recent issues of *Management Science*, *Operations Research*, and *Socio-Economic Planning Sciences* will demonstrate.

The fourth category is concerned with providing analytical tools for dealing with general problem types arising in the above areas. We now outline Wagner's book, which is concerned primarily with this area.

---

[5] See A. K. Sen, *Collective Choice and Social Welfare* (San Francisco, Calif.: Holden-Day, 1971).

[6] See, e.g., the following: H. P. Hatry, "Measuring the Effectiveness of Nondefense Public Programs," *Operations Research*, Vol. 18, No. 5 (September/October 1970), pp. 772–84; E. J. Mishan, "The Postwar Literature on Externalities: An Interpretative Essay," *Journal of Economic Literature*, Vol. 9, No. 1 (March 1971), pp. 1–28.

[7] See C. H. Stevens, "Science, Government, and Citizen Feedback," *Operations Research*, Vol. 18, No. 4 (July/August 1970), pp. 577–91.

[8] See H. Hinrichs and G. Taylor, *Program Budgeting and Benefit Cost Analysis* (Pacific Palisades, Calif.: Goodyear Publishing Co., 1969).

[9] See S. B. Botner, "Four Years of PPBS: An Appraisal," *Public Administration Review*, Vol. 30, No. 4 (July/August 1970), pp. 423–31.

[10] See I. R. Hoos, "Information Systems and Public Planning," *Management Science*, Vol. 17, No. 10 (June 1971), pp. B658–71.

☐ **Wagner: operations research.** This treatise is a comprehensive introduction to the underlying philosophy and fundamental results of operations research. On many fronts it is definitive of the present state of the art. However, as Wagner himself has pointed out,[11] radical departures from the essentially methodological approach of traditional operations research must be developed if continued progress and applications in higher levels of the planning hierarchy are to be achieved.

In any case, this book will provide a useful arsenal of techniques for formulating and solving problems in the public and private sectors for some time to come. The style of the book is informal and designed for a broad range of readers. An excellent bibliography is provided and integrated with the text, including self-study guides for each major area for further in-depth reading. A wide range of theoretical and applied exercises of various degrees of difficulty is given throughout to emphasize problem formulation and solution methods.

On the theoretical side, there is a thorough discussion of linear, non-linear, and dynamic programming. The theory of networks and Markov chains is discussed with associated optimization problems. Fundamental results from queuing theory and dynamic optimization problems are given. Finally, a short introduction to simulation and its uses is presented.

On the applied side, the traditional uses of operations research in production and operational planning serve as major illustrations of the theoretical methods introduced. A wide variety of other applications, however, will convince the reader of the broad scope of applicability of the "operations research approach." Such examples include facilities location and capacity planning, project selection and control, flow and scheduling problems, capital budgeting, and manpower planning.

In summary, this book provides an excellent introduction to the current state of operations research. Though its main focus is on private sector applications, the approach taken requires little imagination to see its applicability to many problems in the regulated industries and the public sector. It is both an excellent text and an invaluable reference to the field.

☐ **de Neufville and Stafford: systems analysis.** Turning now to de Neufville and Stafford, quite a different set of problems is discussed. The authors have attempted to synthesize several approaches to systems analysis including those of microeconomics and operations research. A short description of the main topics covered follows.

After an introductory chapter on the nature of systems analysis, a brief summary of production functions and marginal analysis is given. Several chapters are then devoted to a discussion of fundamental results in mathematical programming and its uses in decision problems. The level of discussion here is fairly elementary, but it is adequate to convey problem formulation and a general understanding of the applicability of mathematical programming in systems analysis.

---

[11] H. M. Wagner, "The ABC's of OR," *Operations Research*, Vol. 19, No. 6 (October 1971), pp. 1259–81.

Following this, the authors return to the realm of applied economics, dealing first with problems of cost estimation and technological forecasting, then with evaluation and selection of projects from a capital budgeting viewpoint. Concepts of individual preference theory and social welfare are then introduced and the problems of traditional cost-benefit analysis are analyzed vis-à-vis distribution effects and externalities. The book concludes with an excellent discussion of the structural and econometric specification and identification of systems models.

It was disappointing that simulation models were not introduced, given their growing importance in systems analysis.[12] A further omission is a discussion of the management environment of systems analysis. In particular, none of the PPBS-related literature is represented.

Because of the broad focus of this book, it is impossible for the authors to deal in depth with any of the major issues it addresses. Nonetheless, I have found this text to be a useful road map for an introductory course on systems analysis and public policy. Moreover, the lucid and nontechnical presentation, and the extensive references at the end of each chapter, should make this book a useful introduction to the field for the non-academician. There are, for the most part, better presentations of the topics covered elsewhere in the literature; but as a first attempt at synthesizing the scattered strands of systems analysis, the authors have performed a valuable and necessary service.

☐ **Systems analysis and management science.** In terms of the framework introduced above, the major deficiency in these two books is their failure to integrate their results with a meaningful planning and control framework. Though few would gainsay the importance of analytical methods in structuring particular problems in the public sector, the experience of the past decade indicates that such studies tend to be useful only when coupled with a reasonable management system. For one thing, it is difficult to understand the context and relative importance of a particular project without an ongoing management control system. Second, data for such studies are usually not readily available until a program structure is developed integrating managerial responsibilities with program outputs.[13] Even when such data are available or can be made available at some cost, the refinement and use of study results in an operational environment cannot proceed without organizational understanding and support. Finally, if a primitive organizational structure and information system are fixed inputs to a systems study, the value of the study may well be negative.

Lest one be too critical, it should be noted that research in systems analysis as it relates to management decision-making is in its infancy. There are some exceptions at the policy level and in particular applications areas, but in many instances it appears that an undisciplined transfer of private sector techniques has been attempted at the middle management and operational levels without adequate thought as to their suitability.[14]

---

[12] See several articles on simulation methodology and use in J. M. Beshers, ed., *Computer Methods in the Analysis of Large-Scale Social Systems* (Cambridge, Mass.: M.I.T. Press, 1968).

[13] For a discussion of program structure and related information structures in the PPBS context, see Hinrichs and Taylor, *op cit.*, Parts I and IV.

Among the issues which are important to resolve here, the following seem particularly pressing. First, how should program and organization structure be related? It has been understood for some time that the nature of program activities may require alternative organizational forms for efficient accomplishment.[15] Little of this understanding has filtered into the public sector. In particular, there is an unfortunate lack of interest in centralization-decentralization issues. We seem to be heading in the direction of duplicating the early errors of the socialist countries by overcentralization on the control side. On the planning end, there seems to be little understanding in practice of aggregation and decomposition effects or of the use of coordination and incentive schemes for promoting decentralized information requirements. When compounded with the eager ignorance of many consulting firms and our lack of understanding of large-scale information systems, the results can be disastrous.

One may hope that current research on these problems will soon be synthesized with the work reviewed here. In the meantime, de Neufville and Stafford and Wagner have presented a readable and innovative summary of the foundations of systems analysis and the methodology of operations research which will serve as important cornerstones of this synthesis.

---

[14] See Hoos, *op. cit.*, for some salient examples in land use information systems.

[15] See P. Lawrence and J. Lorsch, *Organization and Environment* (Boston: Harvard Business School, Division of Research, 1967).

The theory of peak-load pricing: a final note

George W. Wilson

Professor of Economics
Indiana University

This note presents a formal reconciliation of two approaches to the analysis of peak and off-peak pricing. It argues that the differences reside in varying conceptions of the time period over which costs and revenues are calculated. Both versions yield formally correct solutions so that no substantive problems emerge, as noted in the reconciliation. However, varying rates of production, implicit in the two approaches, may violate the cost linearity assumptions made in both and hence vitiate the traditional conclusions.

The theory of peak-load pricing is now rather old hat and, in fact, has been "correctly" resolved for over a decade. That is, under the simplifying assumptions of cost linearity (both short- and long-run) and linear demand functions, a neat, determinate solution can be shown to exist as long as we ignore "second best" problems and income distribution effects. Yet there exist two alternative formulations of the problem whose interrelatedness seems not to have been fully recognized. The purpose of this note is simply to tidy up a corner of an interesting and not totally irrelevant or impractical problem.[1]

The two alternative approaches are those of Steiner and Hirshleifer (S-H) on the one hand compared with Williamson on the other.[2]

Starting with the two-period situation there are two cases. First, for any given plant capacity, including optimal capacity, the demand function at the peak (D_1) intersects the capacity constraint at $P_1 > b > 0$ ($b = SRMC$ of operation), whereas the demand function in the off-peak period does not intersect the capacity constraint at a price $P_2 \geq b$, and $Q_1 > Q_2$. Steiner refers to this as the fixed peak and Wellisz,[3] whose work incidentally is never cited in this connection despite a correct solution that is really something of a precursor to Williamson's, refers to it as the "insignificant off-peak" case.

The two solutions are as follows. For S-H the approach involves maximizing a profit function (π) subject to a zero profit constraint over the peak plus off-peak periods. Thus,

$$\text{Max}\pi = P_1Q_1 + P_2Q_2 - b(Q_1 + Q_2) - B^*Q_1,\ Q_1 > Q_2, \quad (1)$$

subject to $\pi = 0$, where B^* = unit capacity cost for peak output (Q_1). Performing the familiar manipulations we can readily derive the "classical" S-H solutions, namely,

$$P_1 = b + B^* \quad (2)$$

---

[1] This said despite Ralph Turvey's recent cold water douche on these aspects of the theory [4].

[2] See Steiner [3], Hirshleifer [1], and Williamson [6].

[3] In [5].

and

$$P_2 = b\,. \tag{3}$$

Williamson's approach seeks to maximize the sum of producers' and consumers' surplus (W) as follows:

$$\begin{aligned} W = (P_1Q_1 + S_1)W_1 + (P_2Q_2 + S_2)W_2 \\ - bQ_1W_1 - bQ_2W_2 - BQ_1\,, \end{aligned} \tag{4}$$

where P_iQ_i = producer revenue, S_i = consumers' surplus, B = per-unit capacity cost for peak output and W_i refers to the proportion of the total demand cycle over which P_iQ_i pertains, and $W_1 + W_2 = 1$. From (4) we derive

$$\begin{aligned} \frac{\partial W}{\partial Q_1} &= P_1W_1 - bW_1 - B = 0 \\ \therefore\ P_1 &= b + \frac{B}{W_1} \end{aligned} \tag{5}$$

and

$$\begin{aligned} \frac{\partial W}{\partial Q_2} &= P_2W_2 - bW_2 = 0 \\ \therefore\ P_2 &= b\,. \end{aligned} \tag{6}$$

In the general case (i.e., number of periods of equal length = n) we have the following solutions to the fixed peak problem:

S-H: $P_i = b + B^*$, for peak price at optimal Q_i, (7)
$P_j = b$, for off-peak prices, and

Williamson: $P_i = b + B/W_i$, for peak price at optimal Q_i, (8)
$P_j = b$, for off-peak prices.

Equations (7) and (8) are equivalent under two sets of circumstances. First, if $B^* = B$, and using the S-H "additive" approach implying that $W_i = W_j = 1$, then equation (8) is the same as (7). Second, if we use Williamson's "proportionate" approach where $\frac{1}{n} = W_i = W_j$ and $\sum_{i=1}^{n} W_i = 1$, then (8) is the same as (7) if $B = B^*/n$.

In the two-period shifting peak situation where $Q_1 = Q_2$ the S-H solution may be derived as follows:

$$\begin{aligned} &\text{Maximize } \pi = P_1Q_1 + P_2Q_2 - 2Q_1b - B^*Q_1\,, \\ &\text{Subject to } \pi = 0\,. \end{aligned}$$

As before, using familiar techniques, we can deduce optimal prices as follows:

$$P_1 + P_2 = 2b + B^*\,. \tag{9}$$

Williamson's results are obtained by setting $Q_2 = Q_1$ in (4) and differentiating to get:

$$P_1W_1 + P_2W_2 = b + B\,. \tag{10}$$

For the general n-period situation we have

$$\text{S-H:}\quad \sum_{i=1}^{n} P_i = nb + B^* \tag{11}$$

and

$$\text{Williamson:}\quad \sum_{i=1}^{n} P_iW_i = b\sum_{i=1}^{n} W_i + B. \qquad (12)$$

Again (12) is equivalent to (11) as long as $B = B^*$ and $W_i = W_j = 1$. Alternatively (11) is equivalent to (12) as long as $B = B^*/n$, $1/n = W_i = W_j$, and $\sum_{i=1}^{n} W_i = 1$.

The implications of these views, however, go beyond any differences in definition of capacity as claimed by Williamson. The differences reside in the implicit definition of the time period over which total costs and revenues are computed. For example, when S-H compute total revenues (or costs) for the peak or off-peak periods, these are deemed to represent total sales at the price prevailing over the length of the period. Thus total revenues for S-H are merely $P_1Q_1 + P_2Q_2$. Williamson's version, using the entire demand cycle and weighting cost and revenues in proportion to the ratio of the length of time of the peak (or off-peak) period to the total cycle, really implies that at a price P_i, held constant over the period of the peak (or off-peak), total sales are Q_i *per sub unit* of the peak. For example, in his illustration showing total revenues equal to $P_1Q_1W_1 + P_2Q_2W_2$ over an entire 24-hour cycle with $W_1 = \frac{1}{3}$ and $W_2 = \frac{2}{3}$, this must be interpreted to mean that the Q_i are sales *per hour* so that total revenue during the peak is $8P_1Q_1$ and during the off-peak is $16P_2Q_2$. Thus revenue per hour, per cycle, becomes $\frac{8P_1Q_1 + 16P_2Q_2}{24} = \frac{1}{3}P_1Q_1 + \frac{2}{3}P_2Q_2$. In short, S-H interpret Q_i as total sales over the entire period at price P_i whereas Williamson interprets the Q_i as sales per subunit of a subperiod of the total cycle. For S-H therefore capacity B^* is geared to total output over the entire subperiod. Thus Williamson's $B = B^*/n$, where n refers to the number of time units in the subperiod divided by the number of time units over the cycle. As long as the subperiods are equal and thus contain an equal number of time units, the two approaches are made consistent by either of the two interpretations given above. In other words, Williamson's Q_i are sales for a unit of time that is less than the period of the peak or off-peak, whereas the Q_i's for S-H are sales per unit of time equal to the period of the peak or off-peak. B^* for S-H is thus defined over a different and longer time span and is equivalent to nB, à la Williamson, when periods are of equal length. Williamson is not therefore correct when he argues that the S-H approach requires "that cost be redefined for each change in the number of subperiods."[4]

The question at issue is whether it is more appropriate to view "capacity" in terms of a shorter or longer period. Technically there should be no problem because the amount of capacity required to produce Q_1 units over, let us say, an 8-hour (-day, -week, etc.) period with no stipulation as to the *rate* of production per hour, minute, and so on during the period is the same as the capacity needed to produce Q_1 units at an even flow per hour of $Q_1/8$ for 8 hours: in the former case, the firm can choose its rate of production and is free to choose Q_1/n or $Q_1/8$ if it desires; similarly, in the latter case, while somewhat more constrained, the firm can still choose to produce at a rate Q_1/n $(n > 8)$ and meet the condition that the rate must at least be $Q_1/8$ unit per hour.[5] The essential point for the present discussion is that Williamson defines his capacity in terms of Q_1/n whereas S-H define their capacity in terms of Q_1. This necessitates the previously noted equivalence between B and B^* but it does not require in either version a respecification of capacity cost. Williamson's approach in this sense is not more general than that of S-H.

---

[4] See [6], p. 810. His proof of this assertion amounts to imposing his version of rate of sales (Q_i, discussed above) upon the S-H formulation. See [6], p. 822, note 11. The formal equivalence between the two approaches has been noted above and does not require a redefinition of capacity cost for varying subperiods in the S-H interpretation.

However, Williamson's approach is more general in the sense that he can deal with cases where the subperiods are of unequal duration (i.e., where $W_i \neq W_j$). That is, where $W_i \neq W_j$ and $\sum_{i=1}^{n} W_i = 1$, we have the optimal solution as

$$\sum_{i=1}^{n} k_i P_i = n(b + B)\,, \tag{13}$$

where

$$\sum_{i=1}^{n} k_i = n\,.$$

There is no easy equivalence between this and the S-H approach except when $k_i = k_j = 1$, which is simply the equal length situation previously analyzed.

In short, aside from this latter point, the two approaches are parallel and from an algebraic point of view there is nothing to choose between them. On the other hand, the S-H geometrical presentation is, as Williamson notes, "unconventional." Construing the W_i as proportions not only permits a somewhat more intuitive interpretation but renders the geometry more usual. Whether this is a real advantage or not is esentially a matter of taste.

References

1. Hirshleifer, J. "Peak Loads and Efficient Pricing: Comment." *Quarterly Journal of Economics*, Vol. 72, No. 288 (August 1958), pp. 451–62.
2. ———. "The Firm's Cost Function: A Successful Reconstruction." *Journal of Business*, Vol. 35, No. 3 (July 1962), pp. 235–55.
3. Steiner, P. O. "Peak Loads and Efficient Pricing." *Quarterly Journal of Economics*, Vol. 71, No. 285 (November 1957), pp. 585–610.
4. Turvey, R. "Peak-Load Pricing." *Journal of Political Economy*, Vol. 76, No. 1 (January/February 1968), pp. 101–14.
5. Wellisz, S. H. "Regulation of Natural Gas Pipeline Companies: An Economic Analysis." *Journal of Political Economy*, Vol. 71, No. 1 (February 1963), pp. 30–44.
6. Williamson, O. E. "Peak-Load Pricing and Optimal Capacity." *The American Economic Review*, Vol. 56, No. 4 (September 1966), pp. 810–27.

---

[5] Of course, different rates of producing a fixed quantity may seriously challenge the cost linearity assumptions along the lines suggested by Hirshleifer [2]. Others have commented on the lack of realism in the linearity assumption (see [4]), but this is not the issue here.

Note on "gold plating" or "rate base padding"

Edward E. Zajac

Head-Economic Analysis Research Department
Bell Telephone Laboratories

This note analyzes the ancient "gold plating" or "rate base padding" issue. It is shown that padding is profitable for the regulated firm only if it pursues certain expansion paths, and only if it has exhausted all positive-income-producing investment opportunities and must resort to negative-income-producing investments to bring down its rate of return to that allowed.

■ In their seminal paper, Averch and Johnson pointed to the possibility of regulation causing the overintensive substitution of capital for other production factors.[1] The issue they raised has received considerable attention[2] and has been comprehensively reviewed by Baumol and Klevorick.[3] However, surprisingly little attention has been paid to a much older issue of possible overintensive capital use, namely that of "gold plating" or "rate base padding." Although this issue is widely discussed verbally by those involved with regulation, it seems to appear rarely in the regulatory literature, and its economic analysis seems to be absent from the economic literature. This note attempts to remedy that absence.

The rate base padding charge is based on the regulatory formula:

Allowed earnings shall not exceed
the firm's rate base times the
"fair" rate of return.

The charge observes that even with a properly set fair rate of return, this formula allows the firm to increase its allowed earnings by artificially increasing its rate base. With more rate base, the firm can ask for a rate increase to realize the increased allowed earnings. So almost from the beginning of rate-of-return regulation, regulated firms have been accused of "padding the rate base" by adding nonproductive capital. In the past, rate base padding allegations have generally been of two forms: (1) that regulated firms "gold plate," i.e., use unnecessarily expensive materials and designs, or (2) that they maintain excessive spare capacity. Recently, the rate base padding allegation has taken a new form, namely that regulated firms will tend to penetrate into subsidiary markets at nonprofitable prices, not as classic predatory pricers, but primarily to pad the rate base while enjoying the secondary benefit of preempting markets or forestalling entry.[4]

---

[1] See [1].

[2] See Bailey and Malone [2], Kafoglis [6], Takayama [7], and Zajac [8].

[3] See [3].

[4] The market penetration issue seems to have been raised first by Averch and Johnson (see [1], Parts II and III). This has led to some confusion, namely, that the market penetration phenomenon is a special case of the possible capital

In defense of the regulatory formula it has been argued that capital which has been raised to "gold plate" or "pad" is costly; it is silly to incur costs for non-revenue-producing capital when for the same costs productive capital can be obtained. And so the debate has been joined.

It is interesting first to consider the model which seems to underlie verbal discussions of rate base padding. In this model, the production side of the firm is assumed to be summarized by a single factor of production, capital. Imagine then that the firm confronts investment opportunities arranged in decreasing order of attractiveness, and that the firm has worked its way through its investment schedule to the point where K is the magnitude of its total capital investment. Let $I = I(K)$ be the firm's *net* income or earnings, i.e., gross revenues minus operating expenses. Then, if i is the cost of capital, the firm's profit π over and above the cost of capital is given by

$$\pi = \pi(K) = I(K) - iK. \tag{1}$$

As usual in such analyses, we neglect depreciation and definitional distinctions between "fair value," "replacement value," and "original value" rate bases and simply let K be the firm's rate base. Then the regulatory formula or constraint is

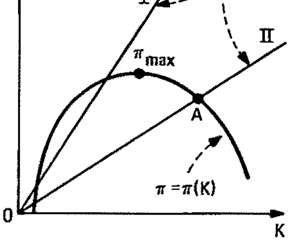
$$I \leq fK, \tag{2}$$

where f is the "fair" rate of return. From equations (1) and (2) an alternative form of the regulatory constraint is

$$\pi \leq (f - i)K. \tag{3}$$

FIGURE 1

INTERSECTIONS OF THE REGULATORY LINE AND THE PROFIT CURVE



If the profit-maximizing firm does not gold plate or rate base pad, its operating point is obvious and depends on where the curve $\pi = \pi(K)$ cuts the straight line $\pi = (f - i)K$. See Figure 1. For an intersection to the left of the point of maximum profit at the top of the profit curve, $\pi = \pi(K)$ (Figure 1, Line I), regulation has no effect; the firm would operate at the top of the profit curve, just as it would in the absence of regulation. For an intersection to the right of $\pi_{\max}$ (Figure 1, Line II), regulation has an effect, and the firm's operating capital would be given by the intersection point A. At this value of capital, the firm's profits are the highest it can attain while satisfying the regulatory constraint. Corresponding to the usual assumption that an expanding firm will work its way through a schedule of decreasingly attractive investment opportunities, we have assumed that $\pi = \pi(K)$ is of a concave, inverted "U" shape. Although the detailed results change when this assumption is violated, the methodology to be presented still applies.

Consider next the case of rate base padding. For simplicity assume that this occurs in the most blatant form where the firm acquires completely nonproductive capital, K^*. Then equations (1) and (3) become

$$\pi = \pi(K,K^*) = I(K) - i(K + K^*) \tag{4}$$

---

oversubstitution for labor pointed out in Averch and Johnson [1], Part I. However, as this note attempts to make clear, market penetration is really an example of rate base padding. See also Johnson [5].

$$\pi \leq (f - i)(K + K^*) \,. \qquad (5)$$

Equation (4) shows that padding tends to *decrease* the firm's profit, while equation (5) shows that padding *increases* the allowed earnings and profit. Thus equations (4) and (5) mirror the argument and counterargument of the debate.

At the outset, it is clear that in the case shown by Figure 1, Line I, where $\pi = \pi(K)$ and $\pi = (f - i)K$ intersect to the left of $\pi_{\max}$, the maximum profit point, padding is not profitable. The best the firm can hope for in any event is its unregulated profit maximum. Since in this case the unregulated profit maximum can be achieved, padding can only decrease profits.

Of principal interest is the case where regulation prevents the attainment of the unregulated profit maximum. Imagine that the firm expands productively to an amount K of capital, and subsequently pads with an amount K^* until it satisfies the regulatory constraint. Padding reduces profits by iK^*. Thus in a diagram of π versus $K_t = K + K^*$, padding is represented by a straight line starting at the point where $K_t = K$ and with negative slope of magnitude i. Figure 2 shows such a diagram, where the firm's income and profit are represented by a solid curve as a function of total capital, $K_t = K + K^*$, and by a dotted curve as a function of K (nonpadding firm). In the case shown in Figure 2, the intersection of $\pi = \pi(K)$ and $\pi = (f - i)K$ occurs between the points of $\pi'(K) = d\pi(K)/dK = 0$ and $\pi'(K) = -i$. The firm is shown padding starting from a point to the left of B, the intersection of the dotted curve and $\pi = (f - i)K$. Because $\pi'(K) > -i$ for all points on $\pi = \pi(K)$ to the left of B, it is clear that Point A, the intersection of the solid curve and $\pi = (f - i)K$, will lie below B. Obviously, in this case padding from a point on $\pi = \pi(K)$ to the right of B would also be of no avail. It would simply reduce profit from that obtainable at B. Thus padding in the case shown in Figure 2 is of no advantage.

FIGURE 2

INTERSECTION BETWEEN $\pi_{\max}$ AND $\pi'(K) = -i$, PADDING IS UNPROFITABLE

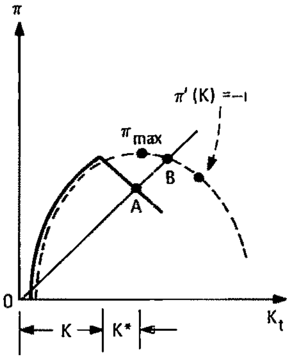
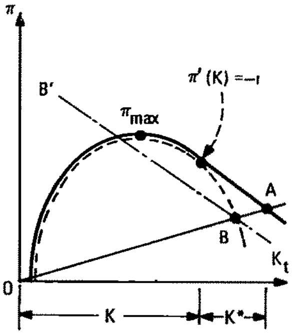


Figure 3, on the other hand, shows the intersection of $\pi = \pi(K)$ and $\pi = (f - i)K$ to the right of the point where $\pi'(K) = -i$. Clearly in this case the firm's best strategy is to expand productively to that point of $\pi'(K) = -i$ and then pad to Point A until the allowed rate of return is achieved. Indeed, in Figure 3 consider the line B′B passing through B and of slope $-i$. Productive expansion to any point on $\pi = \pi(K)$ lying above B′B, followed by padding, would yield a greater profit than that at Point B. However, padding from the $\pi'(K) = -i$ point yields the greatest profit. Note that for a sufficiently flat $\pi(K)$ curve, $\pi'(K) > -i$ everywhere that π is positive. In such a case, the firm will not find padding profitable.

Put in other terms, the firm which decides to pad would clearly desire the highest attainable solid curve. If for positive π there is a point at which $\pi'(K) = -i$, this will be the solid curve shown in Figure 3. It coincides with $\pi = \pi(K)$ from $K = 0$ to the point of $\pi'(K) = -i$; to the right of this point it is above the dotted $\pi = \pi(K)$ curve. If $\pi = (f - i)K$ intersects to the left of $\pi'(K) = -i$, where the highest attainable solid curve does not differ from the dotted curve, padding is of no advantage. If $\pi = (f - i)K$ intersects to the right of $\pi'(K) = -i$, where the solid curve lies above the dotted curve, the firm's better strategy is to pad so as to follow the solid rather than the dotted curve.

FIGURE 3

INTERSECTION TO THE RIGHT OF $\pi'(K) = -i$, PADDING IS PROFITABLE



Consideration of the firm's net income, $I = I(K)$, gives a straightforward economic interpretation of the above analysis. Recall that income, $I(K)$, has been defined as gross revenues minus operating expenses, without regard to capital costs. Thus negative income means that the firm's gross revenues do not even cover operating expenses, much less the cost of capital. Now from equation (1),

$$\pi'(K) = I'(K) - i\,.$$

Hence $\pi'(K) = -i$ implies $I'(K) = 0$; the decision value of K for padding versus no padding corresponds to the point of maximum income. Further expansion of productive capital, K, beyond the decision value means that the firm must use investments which decrease $I(K)$, that is, negative-income-producing investments whose gross revenues do not even cover operating expenses.

Not surprisingly, the analysis has shown that if the regulated firm must resort to negative-income-producing investments to reduce earnings to those that are allowed, it is better for the firm to use *zero-income-producing* investments rather than the negative-income-producing investments. Only in this case is padding profitable. Again, we note that $I'(K)$ may be positive over the entire range of production possibilities which yield positive profits. That is, in this range the firm may always be able to find still another investment opportunity of ΔK which will yield a positive ΔI. In this case, padding will always be unprofitable.

The recent literature bases the theory of the regulated firm not on the simple model given above but on the classic model which assumes that the firm's profit is given by

$$\pi = p(q)q(K,L) - wL - iK\,. \qquad (6)$$

Here L is the labor input and w the unit labor cost, $q(K,L)$ is the production function, and $p(q)$ the inverse demand function. To relate this model to the simpler model discussed so far, one must assume some expansion path for the firm, i.e., some relationship $L = L(K)$ for the labor used as a function of capital. For some expansion paths, it may indeed turn out for positive K and π that $I'(K) > 0$, or equivalently $\pi'(K) > -i$, and hence that padding is unprofitable. An example is the locus of points satisfying

$$\partial\pi/\partial L = [d(pq)/dq]\partial q/\partial L - w = 0\,,$$

assumed by Baumol and Klevorick as the regulated firm's expansion path.[5] Along this path, the usual assumptions $\partial q/\partial K > 0$, $\partial q/\partial L > 0$, $w > 0$ yield $d(pq)/dq > 0$; thus $d\pi/dK = [d(pq)/dq]\partial q/\partial K - i > -i$, and padding does not profit the firm. This is consistent with the proofs given in Edelson and Zajac[6] that padding is unprofitable for the regulated, profit-maximizing firm with perfect knowledge and complete freedom to choose its capital-labor mix.

References

1. AVERCH, H. and JOHNSON, L. L. "Behavior of the Firm Under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–69.

---

[5] See [3].

[6] See [4] and [8], respectively.

2. Bailey, E. E. and Malone, J. C. "Resource Allocation and the Regulated Firm." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 129–42.
3. Baumol, W. J. and Klevorick, A. K. "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–90.
4. Edelson, N. M. "Resource Allocation and the Regulated Firm." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 374–78.
5. Johnson, L. L. "Communications Satellites and Telephone Rates: Problem of Government Regulation (1969)," quoted in A. K. Jones, *Case and Materials on Regulated Industries*, Brooklyn, N.Y.: The Foundation Press, 1967, pp. 340–41.
6. Kafoglis, M. Z. "Output of the Restrained Firm." *The American Economic Review*, Vol. 59, No. 4 (September 1969), pp. 583–89.
7. Takayama, A. "Behavior of the Firm Under Regulatory Constraint." *The American Economic Review*, Vol. 59, No. 3 (June 1969), pp. 255–60.
8. Zajac, E. E. "A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model." *The American Economic Review*, Vol. 60, No. 1 (March 1970), pp. 117–25.

Cash-flow patterns and the choice of a depreciation method

Alan Kraus

Assistant Professor of Finance
Stanford University

and

Ronald J. Huefner

Assistant Professor of Accounting and Management Science
State University of New York at Buffalo

The determination of annual depreciation charges is important in several management contexts. A major context is the determination of book value and net income (and the associated rate of return on investment). This is important for all firms, but especially for those in regulated industries due to the role of income in rate setting.

This paper builds upon the notion of change in present value as the ideal definition of annual depreciation. The usual approach is to develop depreciation methods from cash-flow patterns. This paper, however, proceeds in the opposite direction, beginning with depreciation methods and determining the cash-flow patterns for which each method is appropriate. To achieve this, two generalized depreciation patterns are defined (which include the commonly used methods as special cases). The results are developed mathematically and are illustrated numerically. Consideration is also given to problems of application, particularly in analysis of the choice of depreciation pattern when projected cash flows differ from the patterns consistent with the general depreciation method to be used.

1. Introduction

■ The choice of depreciation schedules for a firm's capital assets can have important implications for financial magagement. Since depreciation determines change in book value and is deducted in calculating net income, the choice of the depreciation method affects the pattern over time of return on investment.[1] In the case of regulated industries, measures of book value, net income, and return on investment are particularly important for rate determination.

We define depreciation as the annual decrease in the value of an asset to its owner. Value, in turn, means the present value of all future cash flows which are due to use of the asset.[2] These cash flows are assumed to be after taxes. The best depreciation method for tax purposes can be selected independently.[3] Upon this selection, the tax payments are included as cash flows relating to the asset, along with revenues and operating costs. The determination of the depreciation method for purposes of income measurement is then based on the after-tax flows.

---

[1] See Solomon [10], pp. 72–73.

[2] We consider only "conventional" investments—a net outlay followed by several consecutive periods of net inflows. We rule out consideration of negative depreciation (appreciation). The economic life over which the asset is to be depreciated is the period during which the asset produces positive net cash flows.

[3] See, for example, Schoomer [9]. Note that if tax depreciation differs from book depreciation, many regulatory agencies require either a flow-through of the

The definition of depreciation used here is related to the questions of income and value measurement mentioned above.[4] In an economic sense, a firm's income for a year is the amount of resources the firm could consume (i.e., distribute to its owners) and still expect to be as well off at the end of the year as at the beginning. The present value of the future net cash flows the firm expects is a measure of how "well off" it is. This leads to the following definition of the economic income of an asset for a year:

Net economic income for the year	equals	Net cash flow during the year	less	Decrease in present value over the year.

In the firm's accounts, net income for the year is, basically, the net cash flow during the year from the asset, less the depreciation charge for the asset for the year. If book income is to represent economic income, we are led to a definition of depreciation as change in present value. Similarly, suppose return on investment for a year is defined as the asset's net income for the year, divided by its value at the beginning of the year. Then return on investment will be uniform over all years of the asset's life if depreciation equals change in present value. For convenience, we will refer to depreciation patterns that satisfy this condition as "present-value depreciation."

Any discussion of a present-value approach involves the problem of choosing a discount rate. There are two main alternatives with respect to the interpretation of this rate. The first approach is to regard it as the rate appropriate for evaluating the firm's investment projects. Under this interpretation, the present value represents the discounted stream of future cash flows expected from the asset. We cannot, in general, expect this present value to equal the asset's acquisition cost. To use this value as the depreciable base for the asset, it would be necessary to establish the asset on the books at its present value. The excess of present value over cost would be considered an unrealized gain.[5] The second interpretation of the discount rate involves the use of acquisition cost as the depreciable base for the asset. With this approach, the present value must equal acquisition cost and so its value is known. Then the discount rate becomes the internal rate of return of the asset and must be calculated from the present value and the expected cash flows. For the purposes of this paper, we assume that the discount rate equals the internal rate of return, so that the initial present value of the asset equals its cost.

This paper considers two classes of depreciation patterns and their relationships to patterns of cash flows. Previous work has considered

---

difference, or normalization of income. For analysis of these effects, see Bierman [3] and Linhart [7].

[4] These definitions of asset value, income, and depreciation have been employed by many economists. See, for example, Fisher [5], Chap. 14, Sec. 6, and Hicks [6], Chap. 14.

[5] The unrealized gain might be shown as a contra-account to the present value, so that the net amount would equal cost. The unrealized gain would be reduced to zero ("realized") over the life of the asset. See, for example, Bierman [2].

the depreciation patterns that are appropriate to particular sets of cash flows. A treatment of the latter problem in the context of a general, multi-asset model is presented in Stauffer.[6] We approach the problem from the opposite direction. That is, given certain depreciation patterns, we determine the shape of the cash-flow stream consistent with the depreciation charge being equal to annual change in present value.[7]

If the choice of depreciation pattern were unconstrained, the firm could simply use the set of changes in present value as its depreciation schedule.[8] At the other extreme, adherence to a single depreciation pattern such as straight-line, perhaps because of tradition or regulatory constraints, may result in large departures from change in present value in the case of many assets. A compromise between these extremes is to allow use of a class of depreciation patterns having some relatively simple shape.

In the present paper, we consider two general classes of depreciation patterns. We show that the use of these two classes, which contain the straight-line and sum-of-the-years'-digits methods as special cases, is consistent with a variety of cash-flow patterns. We also consider the implementation of these results, including the choice of depreciation pattern, when expected cash flows deviate from the patterns implied by any of the methods being considered.

From the point of view of a regulatory agency, our analysis serves to indicate the implicit assumptions that are made about the shape of future cash-flow patterns in the industry when a particular depreciation method is prescribed. If, for example, straight-line depreciation is required, the agency is acting as if it expects future cash flows to decrease linearly. Since the goal of a stable rate of return takes on particular importance in industries subject to rate regulation, the agency should employ this type of analysis in prescribing depreciation methods.

Admittedly, in regulated industries, a particular depreciation method (most often, straight-line) may be specified. In the interest of practicality, we suggest at least broadening the acceptable set of methods to include the two general patterns discussed, rather than limiting it to one or more of the traditionally used methods (straight-line, double-declining-balance, and sum-of-the-years'-digits methods), each of which is a special case of one of the general patterns. Such a broadening would retain the benefits of using systematic depreciation methods, but would substantially improve the matching of costs and revenues and hence the measure of net income. In view of the importance of net income in rate regulation, such a proposal seems worthy of consideration.

---

[6] See [11].

[7] Previous work on this topic has been done by Anton [1] and Reynolds [8] as well as Stauffer [11]. Though the focus of their work was the computation of depreciation patterns from cash-flow patterns, for certain depreciation methods they did indicate the shape of the related cash-flow patterns. In the present paper, however, we derive the results for more general cases and consider also the implementation of the results.

[8] As demonstrated by Stauffer [11], this is a necessary condition for return on investment to equal internal rate of return.

2. Notation, assumptions, and basic relations

■ We employ the following notation:

C_t Net cash flow from the asset in year t; assumed to occur just prior to the end of year t.

V_t Present value at the end of year t of future net cash flows.

D_t Depreciation charge applied to the use of the asset during year t.

d_t Depreciation charge for year t expressed as a fraction of initial present value.

r Discount rate; annual compounding is assumed.

The asset is assumed to be acquired at the beginning of year zero and to yield its final, non-zero cash flow in year T. Thus,

$$V_t = \begin{cases} \sum_{\tau=t+1}^{T} C_\tau/(1+r)^{\tau-t}, & t = 0, 1, \ldots, T-1 \\ 0 & t \geq T. \end{cases} \tag{1}$$

The depreciation charge for a particular year is defined to be the change in the present value of the asset during that year:

$$D_t = V_{t-1} - V_t\,. \tag{2}$$

It follows from the definition of present value that the sum of the present value at the end of year t and the cash flow during year t, discounted one year, equals the present value at the end of the preceding year:

$$(V_t + C_t)/(1+r) = V_{t-1} \quad \text{from (1)}, \tag{3}$$

$$C_t = D_t + rV_{t-1} \quad \text{from (3) and (2)}, \tag{4}$$

$$\sum_{\tau=t}^{T} D_\tau = V_{t-1} \quad \text{from (2) and (1)}, \tag{5}$$

$$C_t = D_t + r\sum_{\tau=t}^{T} D_\tau \quad \text{from (4) and (5)}. \tag{6}$$

Equation (6) provides a relation between a particular pattern of depreciation charges and the cash flow for year t implied by that pattern. In other words, given (1) and a particular set of depreciation charges, for the depreciation charges to be defined as in (2), the cash flows must obey equation (6).

For our purposes, it is convenient to deal with depreciation charges expressed as fractions of the initial present value. It is also desirable to rewrite (6) so that the left-hand side is the ratio of the cash flow in year t to the first year's cash flow:

$$d_t = D_t/V_0\,, \tag{7}$$

$$D_t = d_tC_1/(d_1 + r) \quad \text{from (4) and (7)}, \tag{8}$$

$$C_t/C_1 = (d_t + \sum_{\tau=t}^{T} d_\tau)/(d_1 + r) \quad \text{from (6) and (8)}. \tag{9}$$

Equation (9) is the basic relation for the following sections. Given any set of depreciation charges, it expresses the ratio of the cash flow in year t to the cash flow in the first year that is implied by that set of depreciation charges in order that return on investment be constant over the asset's life. In addition to (9), the following sections will employ the fact that the initial present value is to be depreciated to zero over the useful life of the asset:

$$V_0 = \sum_{t=1}^{T} D_t , \tag{10}$$

$$\sum_{t=1}^{T} d_t = 1 \qquad \text{from (7) and (10)}. \tag{11}$$

3. Constant-absolute-change depreciation

■ We first consider the case of depreciation charges that increase or decrease by a constant amount each year. Varying the increment in annual depreciation provides flexibility in the patterns of cash flows consistent with present-value depreciation. The straight-line and sum-of-the-years'-digits methods are special forms of this class of depreciation patterns.

Expressing depreciation charges as fractions of initial present value, let $\Delta d = d_t - d_{t-1}$ be the constant absolute change in annual depreciation. From the definition of Δd we can express the depreciation for year t in terms of Δd and the first year's depreciation:

$$\begin{aligned} d_t &= d_{t-1} + \Delta d \\ &= d_{t-2} + 2\Delta d \\ &\ \vdots \\ &= d_1 + (t-1)\Delta d . \end{aligned} \tag{12}$$

Summing both sides of (12) and using (11), we can express the first year's depreciation in terms of Δd and thus express the depreciation for year t in terms of Δd:

$$1 = \sum_{t=1}^{T} d_t = T d_1 + (1/2)(T-1)T\Delta d , \tag{13}$$

$$d_1 = [2 - (T-1)T\Delta d]/2T , \tag{14}$$

$$d_t = [(2t - T - 1)T\Delta d + 2]/2T . \tag{15}$$

Substituting (14) and (15) into (9), we can express the cash flow for year t, implied by a particular constant-absolute-change depreciation pattern, in terms of the first year's cash flow and the annual change in depreciation. This expression may be written as follows:

$$\frac{C_t}{C_1} = \frac{[1 + r(T - t + 1)][2 + (t-1)T\Delta d] - (T-t)T\Delta d}{2(1 + rT) - (T-1)T\Delta d} . \tag{16}$$

The shape of the pattern of cash flows consistent with constant-absolute-change depreciation, as given by (16), is analyzed in Appendix 1. It is shown that a negative value of Δd (i.e., decreasing depreciation charges) requires cash flows that decrease in the future by a decreasing amount. For a positive value of Δd, the implied cash-flow pattern may be a decreasing or an increasing one, depending on the size of Δd.

The special case of $\Delta d = 0$ (i.e., constant depreciation charges) corresponds to straight-line depreciation. This depreciation pattern implies cash flows that decrease by a constant amount each year, as was noted by Anton.[9] The sum-of-the-years'-digits method is also

---

[9] In [1].

TABLE 1

CONSTANT–ABSOLUTE–CHANGE DEPRECIATION
REQUIRED CASH FLOWS (C_t/C_1)

ANNUAL CHANGE IN FRACTIONAL DEPRECIATION CHARGE, Δd	CASH FLOW OF YEAR, t			
	2	3	4	5
–0.09	0.74651	0.50348	0.27093	0.04883
–0.08	0.76097	0.53170	0.31219	0.10243
–0.07	0.77692	0.56282	0.35769	0.16153
–0.06	0.79459	0.59729	0.40810	0.22702
–0.05	0.81428	0.63571	0.46428	0.30000
–0.04	0.83636	0.67878	0.52727	0.38181
–0.03	0.86129	0.72741	0.59838	0.47419
–0.02	0.88965	0.78275	0.67931	0.57931
–0.01	0.92222	0.84629	0.77222	0.70000
0	0.96000	0.92000	0.88000	0.84000
0.01	1.00434	1.00652	1.00652	1.00434
0.02	1.05714	1.10952	1.15714	1.20000
0.03	1.12105	1.23421	1.33947	1.43684
0.04	1.20000	1.38823	1.56470	1.72941
0.05	1.30000	1.58333	1.85000	2.10000
0.06	1.43076	1.83846	2.22307	2.58461
0.07	1.60909	2.18636	2.73181	3.24545
0.08	1.86666	2.68888	3.46666	4.20000
0.09	2.27142	3.47857	4.62142	5.70000
0.10	3.00000	4.90000	6.70000	8.40000

ASSET LIFE = 5 YEARS   DISCOUNT RATE = 0.05

a special case of constant-absolute-change depreciation, corresponding to $\Delta d = -2/T(T+1)$.

Actual use of (16) requires the construction of tables of the values of the right-hand side for various values of the parameters. A separate table is required for each combination of values of T and r, giving the cash-flow patterns (i.e., values of C_t/C_1) for various values of Δd. For example, Table 1 is based on an asset life of five years and a discount rate of 5 percent. The use of these tables is illustrated by the following:

EXAMPLE I: Suppose an asset with a five-year life will produce cash flows in years one through five of: \$550.00, \$427.31, \$309.55, \$196.73, and \$88.84, respectively. Assume a discount rate of 5 percent. We divide each cash flow by the first-year amount, \$550.00, and consult Table 1. The reader may verify that the resulting fractions of the first-year flow fit the required pattern for $\Delta d = -0.07$. We proceed to compute the present-value depreciation pattern using the formulas given earlier with $T = 5$, $r = 0.05$, $C_1 = \$550.00$, and $\Delta d = -0.07$:

$$d_1 = \frac{1}{5} - \frac{4}{2}(-0.07) = 0.34 \quad \text{from (14) above,}$$

$$
\begin{aligned}
d_2 &= 0.34 - 0.07 &&= 0.27\,, \\
d_3 &= 0.27 - 0.07 &&= 0.20\,, \\
d_4 &= 0.20 - 0.07 &&= 0.13\,, \\
d_5 &= 0.13 - 0.07 &&= \underline{0.06} \\
& && \ 1.00
\end{aligned}
$$

For this asset, $V_0 =$ \$1410.26. See Table 2.

TABLE 2

CONSTANT–ABSOLUTE–CHANGE DEPRECIATION SCHEDULE FOR EXAMPLE I

YEAR	VALUE AT BEGINNING OF YEAR	CASH FLOW	DEPRECIATION	NET INCOME	RETURN ON INVEST-MENT
1	\$1410.26	\$550.00	\$479.49	\$70.51	5.0%
2	930 77	427 31	380 77	46 54	5 0
3	550 00	309 55	282.05	27.50	5.0
4	267.95	196 73	183 33	13 40	5 0
5	84,62	88.84	84 62	4 22	5.0

4. Constant-proportional-change depreciation

■ A second general depreciation scheme is that in which each year's depreciation charge is a constant multiple of that of the previous year. The proportional change in depreciation from one year to the next may be less than unity (i.e., decreasing charges) or greater than unity (i.e., increasing charges). It is assumed to be positive and different from unity for the derivation of the relations that follow.[10]

Expressing depreciation charges as fractions of initial present value, let $b = (d_t - d_{t-1})/d_{t-1}$ be the constant proportional change in annual depreciation. As noted above, it is necessary to require $b > -1$ and $b \neq 0$. We first express the depreciation for year t, d_t, in terms of b and d_1:

$$
\begin{aligned}
d_t &= (1 + b)d_{t-1} \\
&= (1 + b)^2 d_{t-2} \\
&\ \vdots \\
&= (1 + b)^{t-1} d_1
\end{aligned}
\tag{17}
$$

In a manner similar to the preceding section, we sum both sides of (17) and use (11) to express d_1 in terms of b:

$$
d_1 = b/[(1 + b)^T - 1]\,. \tag{18}
$$

Substituting (17) and (18) into (9) produces, after algebraic manipulation, the following relation for the cash-flow pattern consistent with constant-proportional-change depreciation:

$$
\frac{C_t}{C_1} = \frac{(b - r)(1 + b)^{t-1} + r(1 + b)^T}{b - r + r(1 + b)^T}\,. \tag{19}
$$

It is shown in Appendix 2 that the cash-flow pattern given by (19) when b is negative is one of cash flows that decrease in the future by a decreasing amount. When b is positive but less than r, the implied cash flows decrease in the future by an increasing amount. When b exceeds r, the implied cash flows increase with time by a decreasing amount. For the special case of $b = r$, it can immediately be seen from (19) that $C_t/C_1 = 1$ for all values of t. In other words, this is the case of an asset that produces identical cash flows for all years of its life.

---

[10] The case of unity-proportional-change corresponds to the straight-line case which has been treated previously by others.

In order to make use of (19), it is necessary to construct tables of the values of the right-hand side of (19) similar to the tables of (16) described above. To illustrate the patterns of cash flows implied by constant-proportional-change depreciation, Table 3 has been constructed for the same values of asset life and discount rate as were used for Table 1. Table 3 shows cash-flow patterns consistent with various values of b.

It was mentioned in the previous section that constant-absolute-change depreciation includes the straight-line and sum-of-the-years'-digits methods as particular cases. A similar, but weaker, point may be noted about constant-proportional-change depreciation. The double-declining-balance method involves depreciation charges that decrease by a constant percentage each year. Using our notation,

TABLE 3

CONSTANT–PROPORTIONAL–CHANGE DEPRECIATION
REQUIRED CASH FLOWS (C_t/C_1)

		CASH FLOW OF YEAR, t			
		2	3	4	5
ANNUAL FRACTIONAL CHANGE IN DEPRECIATION, b	−0.50	0.49828	0.24743	0.12200	0.05929
	−0.45	0.54772	0.29897	0.16215	0.08691
	−0.40	0.59580	0.35329	0.20778	0.12048
	−0.35	0.64484	0.41400	0.26394	0.16641
	−0.30	0.69261	0.47745	0.32683	0.22140
	−0.25	0.73970	0.54448	0.39806	0.28825
	−0.20	0.78597	0.61475	0.47777	0.36819
	−0.15	0.83128	0.68787	0.56598	0.46236
	−0.10	0.87549	0.76343	0.66258	0.57182
	−0.05	0.91844	0.84097	0.76737	0.69745
	0.03	0.98419	0.96791	0.95114	0.93387
	0.05	1.00000	1.00000	1.00000	1.00000
	0.10	1.03830	1.08044	1.12679	1.17778
	0.15	1.07478	1.16079	1.25970	1.37344
	0.20	1.10932	1.24051	1.39793	1.58684
	0.25	1.14180	1.31906	1.54064	1.81761
	0.30	1.17215	1.39596	1.68691	2.06514
	0.35	1.20030	1.47071	1.83577	2.32859
	0.40	1.22620	1.54288	1.98624	2.60694
	0.45	1.24983	1.61208	2.13735	2.89899
	0.50	1.27118	1.67819	2.28813	3.20338

ASSET LIFE = 5 YEARS DISCOUNT RATE = 0.05

the fact that the double-declining-balance method uses twice the straight-line rate implies $b = -2/T$. However, the depreciation for the first year under the double-declining-balance method is not given by (18). This is due to the fact that the depreciation charges (the d_t) do not sum to one under the usual definition of double-declining-balance depreciation. Thus, although modifications must be made, the depreciation pattern of the double-declining-balance method is essentially a particular case of constant-proportional-change depreciation.

5. Choice of a depreciation pattern for given cash flows

■ For many investments, predicted future cash flows will not fit closely any of the patterns implied by constant-absolute-change or constant-proportional-change depreciation. In such cases, the choice of the best depreciation method requires further analysis.

We will consider two alternative criteria for judging the "goodness" of a depreciation pattern: the closeness of book income to economic income, and the closeness of book value to present value. Both criteria have particular relevance in the case of a regulated firm. It is desirable that book value of assets over time approximates the present value over time of remaining future cash flows and that book income approximates economic income (cash flow less change in present value).[11]

We are dealing here with predicted cash flows that do not correspond exactly to any of the patterns implied by the systematic depreciation methods described previously. Therefore, depreciation charge is no longer required to equal change in present value. Since book value for each year, as determined by the depreciation pattern, is not necessarily equal to present value for that year, separate symbols are required. We retain V_t as notation for present value and add

B_t = book value of the asset at the end of year t.

Relation (2) above is now replaced by

$$D_t = B_{t-1} - B_t . \qquad (20)$$

Assuming, as before, that the asset is fully depreciated by the end of year T (so that $B_t = 0$ for $t = T, T+1, \ldots$), book value equals the sum of future depreciation charges:

$$\begin{aligned} B_t &= \sum_{\tau=t+1}^{T} D_\tau \\ &= V_0 \sum_{\tau=t+1}^{T} d_\tau \quad \text{from (7)} . \qquad (21) \end{aligned}$$

We retain the assumption that the asset's initial present value equals its acquisition cost:

$$B_0 = V_0 . \qquad (22)$$

---

[11] Consideration was given to another objective: that return on investment for the asset over its life be uniform and equal to the internal rate of return. Derivations analogous to those given below are extremely cumbersome for this objective and are not included here. It should be noted that a depreciation pattern that is "good" in terms of the criteria described above for book income and book value will also tend to be "good" in terms of the criterion for return on investment.

For either criterion—closeness of book income to economic income or closeness of book value to present value—we will assume that the objective is operationally specified as minimizing the sum of squared differences between the quantities in question. For constant-absolute-change depreciation, it is possible to derive analytical expressions for meeting the objective of minimizing the sum of squared differences. These expressions are presented below. As noted later, comparable analytical expressions for the case of constant-proportional-change depreciation cannot be obtained. Therefore, the remainder of this section is concerned only with constant-absolute-change depreciation.

For a given value of T, the sequence of depreciation charges expressed as fractions of initial present value are determined by the value of the parameter Δd, as shown in (15) above. In the following development, we let primes denote differentiation with respect to Δd. From (15),

$$d_t' = (2t - T - 1)/2\,, \tag{23}$$

$$d_t'' = 0\,. \tag{24}$$

We deal first with the objective of minimizing the sum of squared differences between book income and economic income:

$$\begin{aligned} \text{Book income for year } t &= C_t - (B_{t-1} - B_t) \\ &= C_t - V_0 d_t \qquad \text{from (20) and (7).} \\ \text{Economic income for year } t &= C_t - (V_{t-1} - V_t) \\ &= rV_{t-1} \qquad \text{from (3).} \end{aligned}$$

We define the sum of squared differences to be minimized as

$$X_1 = \sum_{t=1}^{T} (C_t - V_0 d_t - rV_{t-1})^2\,. \tag{25}$$

Minimizing X_1 with respect to Δd requires $X_1' = 0$ and $X_1'' \geq 0$. From (25), these conditions can be written as

$$V_0 \sum_{t=1}^{T} d_t d_t' + r \sum_{t=1}^{T} V_{t-1} d_t' - \sum_{t=1}^{T} C_t d_t' = 0\,, \tag{26}$$

$$V_0 \sum_{t=1}^{T} (d_t')^2 + V_0 \sum_{t=1}^{T} d_t d_t'' + r \sum_{t=1}^{T} V_{t-1} d_t'' - \sum_{t=1}^{T} C_t d_t'' \geq 0\,. \tag{27}$$

From (24), it can be seen that condition (27) is always satisfied; the solution to (26) will yield the desired minimization of X_1. Using (15) and (23), one can solve (26) for the value of Δd that minimizes X_1. Letting Δd^* denote the value of Δd satisfying (26), algebraic manipulation of (26) produces the following relation:

$$\Delta d^* = \frac{6 \sum_{t=1}^{T} (T^2 + T - 2t)(rV_{t-1} - C_t)}{T(T - 1)(T + 1)V_0}\,. \tag{28}$$

Operationally, the above relation has the disadvantage of requiring calculation of the sequence of present values, V_t, as of the end of each year of the asset's life. This drawback can be avoided by making use of the following relations derived from (1):[12]

$$\sum_{t=1}^{T-1} V_t = \sum_{t=1}^{T-1} \left[\frac{(1+r)^t - 1}{r(1+r)^t} \right] C_{t+1}, \tag{29}$$

$$\sum_{t=1}^{T-1} tV_t = \sum_{t=1}^{T-1} \left[\frac{(rt-1)(1+r)^t + 1}{r^2(1+r)^t} \right] C_{t+1}. \tag{30}$$

Using (29) and (30) in (28) and simplifying yields

$$\Delta d^* = \frac{12 \sum_{t=1}^{T} (C_t/V_0) - 6r(T+1) - 12}{rT(T-1)(T+1)}. \tag{31}$$

The only information required for calculating Δd^* that is peculiar to the asset being considered is the sum of the predicted future cash flows for the asset divided by its initial value. The value of Δd^* calculated from (31) determines a pattern of depreciation charges. Use of this depreciation pattern over the life of the asset will yield a sequence of annual book income values having smaller mean-squared deviation from corresponding economic income than any other constant-absolute-change depreciation pattern. Use of (31) is illustrated in an example later in the section.

The second objective we consider is that of minimizing the sum of squared differences between book value and present value. We define the sum to be minimized, with respect to Δd, as

$$\begin{aligned} X_2 &= \sum_{t=1}^{T} (B_t - V_t)^2 \\ &= \sum_{t=1}^{T} \Big(V_0 \sum_{\iota=t+1}^{T} d_\iota - V_t\Big)^2 \qquad \text{from (21)}. \end{aligned} \tag{32}$$

The conditions for minimization of X_2, $X_2' = 0$ and $X_2'' \geq 0$, can be written as

$$V_0 \sum_{t=1}^{T} \sum_{\iota=t+1}^{T} d_\iota \sum_{j=t+1}^{T} d_j' - \sum_{t=1}^{T} V_t \sum_{j=t+1}^{T} d_j' = 0, \tag{33}$$

$$\begin{aligned} V_0 \sum_{t=1}^{T} \Big(\sum_{j=t+1}^{T} d_j' \Big)^2 + V_0 \sum_{t=1}^{T} \sum_{\iota=t+1}^{T} d_\iota \sum_{j=t+1}^{T} d_j'' \\ - \sum_{t=1}^{T} V_t \sum_{j=t+1}^{T} d_j'' \geq 0. \end{aligned} \tag{34}$$

From (24), condition (34) is always satisfied, so that minimization of X_2 depends only on satisfying condition (33).[13] Let Δd^{**} denote the value of Δd that satisfies (33). Using (15) and (23), we obtain from (33)

$$\Delta d^{**} = \frac{60 \sum_{t=1}^{T-1} (T-t)tV_t}{T(T^4-1)V_0} - \frac{5}{T^2+1}. \tag{35}$$

---

[12] In deriving (30), use is made of the following general relation:

$$\sum_{i=1}^{n} \frac{i}{(1+r)^i} = \frac{(1+r)^{n+1} - (n+1)r - 1}{r^2(1+r)^n}, \quad \text{for } r > -1.$$

[13] As mentioned earlier, explicit solutions of (26) and (33) for the case of constant-proportional-change depreciation are not obtainable. From (17) and

As was true of (28), it is possible to rewrite the above relation to avoid having to calculate the series of present values V_t. To do this, we make use of (30) and[14]

$$\sum_{t=1}^{T-1} t^2 V_t = \sum_{t=1}^{T-1} \left[\frac{(1+r)^{t+1} + (rt-1)^2(1+r)^t - r - 2}{r^3(1+r)^t} \right] C_{t+1} . \quad (36)$$

Using (30) and (36) in (35) and simplifying yields

$$\Delta d^{**} = \frac{60}{r^3 T(T^4-1)} \times \sum_{t=1}^{T-1} \left\{ (rT - rt + 2)rt - (rT + r + 2)\left[1 - \frac{1}{(1+r)^t} \right] \right\} \times \frac{C_{t+1}}{V_0} - \frac{5}{T^2+1} . \quad (37)$$

Although the above relation appears more cumbersome than (35), it has the distinct advantage of permitting Δd^{**} to be calculated directly from the predicted cash flows. Use of the depreciation pattern determined by the resulting value of Δd^{**} will yield a sequence of book values having smaller mean-squared deviation from corresponding present value than any other constant-absolute-change depreciation pattern. Use of (37) is facilitated by noting that it lends itself to the construction of tables. For given values of r and T, (37) is of the form

$$\Delta d^{**} = a_0 + \sum_{t=1}^{T-1} a_t(C_{t+1}/V_0) , \quad (38)$$

where

$$a_0 = \frac{-5}{T^2+1} , \quad (39)$$

and

$$a_t = \frac{60}{r^3 T(T^4-1)} \times \left\{ (rT - t + 2)rt - (rT + r + 2)\left[1 - \frac{1}{(1+r)^t} \right] \right\} . \quad (40)$$

---

(18), expressions can be obtained for d_t and its derivative with respect to b (the parameter for constant-proportional-change depreciation analogous to Δd). These expressions can be substituted into (26) and (33). However, the relations that result involve weighted sums of powers of b in which the weights are functions of the cash flows C_t [in the case of (26)] or the present values V_t [in the case of (33)]. In general, one cannot transform such a relation into a relation in which b is an explicit function of the other variables (r, T, and C_t and the V_t). For a specific set of data, of course, the values of b satisfying (26) or (33) (i.e., analogous to Δd^* or Δd^{**}) could be found by some successive approximation technique.

[14] In deriving (36), use is made of the following general relation:

$$\sum_{i=1}^{n} \frac{i^2}{(1+r)^i} = \frac{(2+r)(1+r)^{n+1} - (2n+3)r - (n+1)^2 r^2 - 2}{r^3(1+r)^n} , \quad \text{for } r > -1 .$$

TABLE 4

CALCULATION OF Δd^* FOR EXAMPLE II

t	C_t/V_0
1	0.29333
2	0.27333
3	0.24000
4	0.19333
5	0.13333

$$\sum_{t=1}^{T} (C_t/V_0) = 1.13332$$

$$\Delta d^* = \frac{(12)(1.13332)-(6)(0.05)(6)-12}{(0.05)(5)(4)(6)}$$

$$= -0.03336$$

For a given value of T, a table can be constructed in which the entries are values of $a_1, a_2, \ldots, a_{T-1}$, corresponding to various values of r. (The value of a_0 depends only on the value of T.) For example, the entries for $T = 5$, $r = 0.05$ (the values used in Tables 1 and 3) are:

$$a_0 = -0.19231\,,$$
$$a_1 = 0.07328\,,$$
$$a_2 = 0.17949\,,$$
$$a_3 = 0.28119\,,$$
$$a_4 = 0.34071\,.$$

The following example illustrates the calculation and use of Δd^* and Δd^{**}. By design, the cash flows chosen for the example are of a pattern different from any of the patterns implied by constant-absolute-change depreciation.[15]

EXAMPLE II: An asset having a five-year life is purchased for \$1,500. The project cash flows for years one through five are \$440, \$410, \$360, \$290, and \$200, respectively. The yield of this investment is 5 percent. We first calculate Δd^* for this asset by use of (31). See Table 4. Next, we calculate Δd^{**} from (38), using the values of a_0, a_1, a_2, a_3, and a_4 given above. See Table 5. Let D_t^* and D_t^{**} be the depreciation charges for year t calculated by using Δd^* and Δd^{**}, respectively, in (15) and (7). The calculated depreciation charges are found in Table 6. The actual sequences of present value, economic income, and return on investment by year for this asset are found in Table 7.[16] For comparison with the values shown in Table 7, analogous values are calculated for depreciation using, respectively, Δd^*, Δd^{**}, straight-line, and sum-of-the-years'-digits. See Table 8.

TABLE 5

CALCULATION OF Δd^{**} FOR EXAMPLE II

t	a_t	C_{t+1}/V_0	$a_t(C_{t+1}/V_0)$
1	0.07328	0.27333	0.02003
2	0.17949	0.24000	0.04308
3	0.28119	0.19333	0.05436
4	0.34071	0.13333	0.04543

$$\sum_{t=1}^{T-1} a_t(C_{t+1})/V_0 = 0.16290$$

$$\Delta d^{**} = -0.19231 + 0.16290$$

$$= -0.02941$$

TABLE 6

CALCULATED DEPRECIATION CHARGES FOR EXAMPLE II

t	D_t^*	D_t^{**}
1	\$400	\$388
2	350	344
3	300	300
4	250	256
5	200	212

It is clear in the above example that a depreciation pattern based on either Δd^* or Δd^{**} yields results for book value, net income, and

---

[15] The cash flows in the example decrease by an increasing amount. As shown in Appendix 1, cash flows consistent with constant-absolute-change depreciation for $\Delta d < 0$ decrease by a decreasing amount.

[16] Some values in Table 7 differ slightly from the correct value of 5.0 percent for each year due to rounding of dollar quantities.

TABLE 7

ACTUAL PRESENT VALUE, ECONOMIC INCOME, AND RETURN ON INVESTMENT BY YEAR (EXAMPLE II)

YEAR	PRESENT VALUE AT BEGINNING OF YEAR	CHANGE IN VALUE TO END OF YEAR	NET ECONOMIC INCOME	RETURN ON INVESTMENT
1	$1500	$368	$72	4.8%
2	1132	353	57	5.0
3	779	321	39	5.0
4	458	267	23	5.0
5	191	191	9	4.7

return on investment much closer to the corresponding economic values than does either straight-line depreciation or sum-of-the-years' digits depreciation.[17] A more complex depreciation scheme might produce an even better correspondence to economic values than the results shown in Table 8, frames (a) and (b), but this would violate the spirit of our approach. Our basic contention is that it is worthwhile to take account of the projected cash-flow pattern in choosing a depreciation method even if the choice is limited to a fairly simple class of depreciation patterns (e.g., constant-absolute-change depreciation).

6. Multiple projects

■ Our analysis has focused on the evaluation of a single asset. The applicability of this approach, however, must be viewed in terms of its effect when multiple investments occur over several periods of time. The results depend upon the nature of the steady-state conditions which prevail. Davidson[18] has shown that an accelerated depreciation method applied to individual assets stabilizes to a straight-line method in the aggregate in situations where there is maintanence of asset investment over time.[19] This result occurs where there is a uniform annual replacement policy and where the total dollar value of asset investment remains constant. However, this situation seldom occurs in practice. If investment grows, either by inflation (replacements cost more than items replaced) or physical growth, or both, or if replacements occur irregularly, then the aggregate depreciation pattern will be dependent upon the method applied to individual assets.[20]

Since most reporting is done on an aggregate basis, attention should be given to the resultant aggregate depreciation pattern in choosing the method(s) to apply to individual assets. To do this, both future cash flows and future asset transactions must be considered. Simulation models may be a useful device for this purpose. It may be determined, as a result of such analysis, that one needs to select a "wrong" method to apply to individual assets in order that the correct aggregate result be obtained. Also, from the regulatory agency's point of view, it is likely that one, or at most a small set, of

---

[17] The comparison would be even more extreme for an asset having cash flows that increase into the future.

[18] In [4].

[19] This may be generalized to apply to *any* depreciation method.

[20] In this connection, see Solomon [10].

TABLE 8

VARIATIONS IN BOOK VALUE, NET INCOME, AND RETURN ON INVESTMENT BY YEAR RESULTING FROM DIFFERENT DEPRECIATION ASSUMPTIONS (EXAMPLE II)

YEAR	BOOK VALUE AT BEGINNING OF YEAR	DEPRECIATION	NET INCOME	RETURN ON INVESTMENT
(a) DEPRECIATION BASED ON $\Delta d^* = -0.03336$.				
1	\$1500	\$400	\$40	2.7%
2	1100	350	60	5.5
3	750	300	60	8.0
4	450	250	40	8.9
5	200	200	0	0.0
(b) DEPRECIATION BASED ON $\Delta d^{**} = -0.02941$				
1	\$1500	\$388	\$52	3.5%
2	1112	344	66	5.9
3	768	300	60	7.8
4	468	256	34	7.3
5	212	212	−12	–
(c) STRAIGHT–LINE DEPRECIATION.				
1	\$1500	\$300	\$140	9.3%
2	1200	300	110	9.2
3	900	300	60	6.7
4	600	300	−10	–
5	300	300	−100	–
(d) SUM–OF–THE–YEARS'–DIGITS DEPRECIATION.				
1	\$1500	\$500	\$–60	–
2	1000	400	10	1.0%
3	600	300	60	10.0
4	300	200	90	30.0
5	100	100	100	100.0

the various possible methods will be prescribed, hopefully to produce the correct aggregate results for the industry as a whole.

7. Conclusion

■ The approach which we have suggested permits the retention of a set of systematic, easy-to-apply depreciation procedures. The set is broadened to the extent of allowing multiple rates of proportional and absolute change (instead of only double-declining-balance and sum-of-the-years'-digits). This increases the ability to match depreciation charges to expected benefits, resulting in a better measure of income. Perfect correspondence of depreciation charges to change in present value will not in general be attained. Such attainment is possible but not practical, due to the infinite number of depreciation methods required and to the inability to estimate cash flows correctly. The latter would result in deviations or necessitate the continual revision of the depreciation method. The procedure suggested in this paper offers an approach that will yield improved results and still remain within the limits of practicability.

Appendix 1

The constant-absolute-change case

■ The necessary pattern of cash flows for the constant-absolute-change case has been shown to be

$$\frac{C_t}{C_1} = \frac{[1 + r(T - t + 1)][2 + (t - 1)T\Delta d] - (T - t)T\Delta d}{2(1 + rT) - (T - 1)T\Delta d}, \qquad t = 1, 2, \ldots, T.$$

To investigate the shape of the cash-flow pattern expressed above, we write

$$C_{t-1} - C_t = \frac{2r + [r(2t - T - 3) - 2]T\Delta d}{2(1 + rT) - (T - 1)T\Delta d}. \tag{41}$$

We wish to consider the sign of the right-hand side of (41). We first establish limits on the magnitude of Δd. To do so, we make use of the assumption that all depreciation charges are positive—that the asset's present value decreases from each period to the next. Thus, we assume

$$d_t > 0, \qquad t = 1, 2, \ldots, T.$$

We also assume $T > 1$, since the case of one-year life is trivial. From (12) and (14),

$$d_t = \frac{1}{T} - \frac{T - 1}{2}\Delta d + (t - 1)\Delta d$$

$$\Delta d = \frac{-2}{2t - T - 1}\left(d_t - \frac{1}{T}\right), \qquad t = 1, 2, \ldots, T.$$

For $t = 1$,

$$\Delta d = \frac{2}{T - 1}\left(d_1 - \frac{1}{T}\right) < \frac{2}{(T - 1)T} \qquad (\text{since } d_1 > 0).$$

For $t = T$,

$$\Delta d = \frac{2}{T - 1}\left(d_T - \frac{1}{T}\right) > \frac{-2}{(T - 1)T} \qquad (\text{since } d_T > 0).$$

Thus we have established

$$\frac{-2}{(T - 1)T} < \Delta d < \frac{2}{(T - 1)T}. \tag{42}$$

We now return to consideration of the right-hand side of (41). First, we use (42) to show that the denominator is always positive. Thus,

$$2(1 + rT) - (T - 1)T\Delta d > 2(1 + rT) - \frac{2(T - 1)T}{(T - 1)T} = 2rT > 0. \tag{43}$$

The sign of the expression under consideration is therefore determined by the sign of the numerator. We first examine the case of $\Delta d < 0$. Since $t \leq T$,

$$r(2t - T - 3) - 2 \leq r(T - 3) - 2 = r(T - 1) - 2(1 + r)$$

$$\begin{aligned} 2r + [r(2t - T - 3) - 2]T\Delta d & \\ &> 2r + [r(T - 1) - 2(1 + r)]\left[\frac{-2T}{(T - 1)T}\right] \quad \text{from (42)} \\ &= \frac{4(1 + r)}{T - 1} > 0 . \end{aligned}$$

This shows that $C_{t-1} > C_t$ (decreasing cash flows) when $\Delta d < 0$ (decreasing depreciation charges). That the cash flows decrease by a decreasing amount in this case may be seen by using (41) to write

$$(C_{t-2} - C_{t-1}) - (C_{t-1} - C_t) = \frac{-2rT\Delta d}{2(1 + rT) - (T - 1)T\Delta d} C_1 .$$

The left-hand side is the second difference in the cash flows—the difference between two successive differences. We have shown in (43) that the denominator of the right-hand side is always positive. Since $\Delta d < 0$, the whole expression must be positive. This establishes the desired result.

The case of $\Delta d > 0$ is more complicated. If r is sufficiently small, the general situation is as follows:

(1) For very small, positive values of Δd, the cash flows decrease each year, as is true when Δd is negative, but by an increasing amount.
(2) For somewhat larger values of Δd, the cash flows first increase then decrease.
(3) For sufficiently large values of Δd, the cash flows increase every year by a decreasing amount.

If r is above certain limits, either case (3) or both cases (2) and (3) are eliminated. Suppose we view the cash-flow line for the straight-line case as an axis. Relative to this axis, all cash-flow patterns for $\Delta d > 0$ have the same general shape—increasing by a decreasing amount—regardless of the values of Δd and r. The cash-flow patterns for $\Delta d < 0$ still decrease by a decreasing amount with respect to the new axis.

In the special case of sum-of-the-years'-digits depreciation, we are given that

$$d_t = \frac{T - t + 1}{1 + 2 + \cdots + T} = \frac{2(T - t + 1)}{T(T + 1)}, \quad t = 1, 2, \ldots, T ,$$

so,

$$\Delta d = \frac{2(T - t + 1)}{T(T + 1)} - \frac{2(T - t + 2)}{T(T + 1)} = \frac{-2}{T(T + 1)} .$$

Substituting this value of Δd into (16) gives the following necessary pattern of cash flows for the sum-of-the-years'-digits case:

$$\frac{C_t}{C_1} = \frac{[1 + r(T - t + 1)]\left[2 - \frac{2(t-1)}{T+1}\right] + \frac{2(T-t)}{T+1}}{2(1 + rT) + \frac{2(T-1)}{T+1}}$$

$$= \frac{(T - t + 1)(rT - rt + 2r + 2)}{T(rT + r + 2)}, \quad t = 1, 2, \ldots, T.$$

Appendix 2

The constant-proportional-change case

■ The necessary pattern of cash flows for the constant-proportional-change case has been shown to be

$$\frac{C_t}{C_1} = \frac{(b - r)(1 + b)^{t-1} + r(1 + b)^T}{b - r + r(1 + b)^T}, \quad t = 1, 2, \ldots, T.$$

As in Appendix 1, we consider the first differences in the cash flows:

$$C_{t-1} - C_t = \frac{b(r - b)(1 + b)^{t-2}}{b - r + r(1 + b)^T} C_1. \tag{44}$$

In analyzing the sign of (44), we assume, as in Appendix 1, that all depreciation charges are positive. From (17), this requires that

$$b > -1.$$

We now consider the sign of the right-hand side of (44).

(i) Suppose $-1 < b < 0$. Then both the numerator and denominator are negative. Hence $C_{t-1} - C_t > 0$.

(ii) Suppose $0 < b < r$. Then both the numerator and denominator are positive. Hence, again, $C_{t-1} - C_t > 0$.

(iii) Suppose $b = r$. Then the numerator equals zero (and the denominator is positive). Hence $C_{t-1} - C_t = 0$.

(iv) Suppose $b > r$. Then the numerator is negative and the denominator is positive. Hence $C_{t-1} - C_t < 0$.

We next consider the second differences in the cash flows. From (44),

$$(C_{t-2} - C_{t-1}) - (C_{t-1} - C_t) = \frac{b^2(b - r)(1 + b)^{t-3}}{b - r + r(1 + b)^T} C_1.$$

We wish to determine the sign of the right-hand side for each of the four cases employed above.

(i) Suppose $-1 < b < 0$. Then both the numerator and denominator are negative. Hence

$$(C_{t-2} - C_{t-1}) - (C_{t-1} - C_t) > 0.$$

(ii) Suppose $0 < b < r$. Then the numerator is negative and the denominator is positive. Hence

$$(C_{t-2} - C_{t-1}) - (C_{t-1} - C_t) < 0.$$

(iii) Suppose $b = r$. Then the numerator equals zero and the denominator is positive. So $(C_{t-2} - C_{t-1}) - (C_{t-1} - C_t) = 0$.

(iv) Suppose $b > r$. Then both the numerator and denominator are positive. Hence $(C_{t-2} - C_{t-1}) - (C_{t-1} - C_t) > 0$.

Combining the above results, we have established that the constant-proportional-change case implies the following patterns of cash flows:

Decreasing by a decreasing amount when $-1 < b < 0$.
Decreasing by an increasing amount when $0 < b < r$.
Equal for all years when $b = r$.
Increasing by an increasing amount when $b > r$.

References

1. ANTON, H. R. "Depreciation, Cost Allocation and Investment Decisions." *Accounting Research*, Vol. 7, No. 2 (April 1956), pp. 117–34.
2. BIERMAN, H., JR. *Financial Accounting Theory*. New York: Macmillan Co., 1965.
3. ———. "Accelerated Depreciation and Rate Regulation." *The Accounting Review*, Vol. 44, No. 1 (January 1969), pp. 65–78.
4. DAVIDSON, S. "Accelerated Depreciation and the Allocation of Income Taxes." *The Accounting Review*, Vol. 33, No 2 (April 1958), pp. 172–80.
5. FISHER, I. *The Nature of Capital and Income*. New York: Macmillan Co.. 1906.
6. HICKS, J. R. *Value and Capital*. London: Oxford Univ. Press, 1939.
7. LINHART, P. B. "Some Analytical Results on Tax Depreciation." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 82–112.
8. REYNOLDS, I. N. "An Analysis of Depreciation Methods and Bases." Research Paper 9, School of Business Administration, University of North Carolina, Chapel Hill, 1962.
9. SCHOOMER, A. B., JR "Optimal Depreciation Strategies for Income Tax Purposes." *Management Science*, Vol. 12, No. 12 (August 1966), pp. B552–79.
10 SOLOMON, E. "Alternative Rate of Return Concepts and Their Implications for Utility Regulation" *The Bell Journal of Economics and Management Science*, Vol. 1, No 1 (Spring 1970), pp. 65–81.
11. STAUFFER, T. R. "The Measurement of Corporate Rates of Return: A Generalized Formulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 2 (Autumn 1971), pp. 434–69.

Metering costs and marginal cost pricing in public utilities

G. Franklin Mathewson

Assistant Professor of Economics
Faculty of Management Studies
University of Toronto

and

G. David Quirin

Professor of Economics
Faculty of Management Studies
University of Toronto

This paper examines a welfare criterion for marginal cost pricing in the presence of metering costs. The good examined is local telephone service, which is viewed as two separate commodities: a connection to the system, and a charge for the telephone services actually used. The empirical results suggest that the rate structures of North American telephone utilities appear more consistent with rational pricing in the face of metering costs than with Averch-Johnson behavior.

■ Arguments that marginal cost pricing leads to Pareto-optimum resource allocation in a static economy usually do not give explicit consideration to the problem of consumption measurement. Given economists' predilection for ignoring transaction costs, this is scarcely surprising. Yet the industries for which marginal cost pricing is most frequently advocated as a basis for regulation, i.e., public utilities, have technologies such that consumption metering is required if any but the crudest of consumption measures is to be applied. Costs of metering, if undertaken, may constitute a significant fraction of total costs of service, so that criteria for deciding when and what to meter are of more than academic importance. Within the utility field on this continent, metering is almost universally practiced by gas and electric service and for long-distance telephone service, while flat rates are frequently used for water supply and are the predominant rate form found in local telephone service, though prominent exceptions exist. Flat rates permit binary measurement of consumption, while metering permits charges to be based on quantity and time of use. Most multi-part rate structures reflect several dimensions of consumption via demand and commodity charges or more complex structures that attempt to take into account contributions to the peak system load.

For analytical purposes, it is convenient to consider the load attributes embraced within the rate structure as separate commodities for which demand curves can be distinguished. Under a demand-commodity charge tariff, the demand charge can be viewed as the charge for providing a connection to the system and the capacity required to provide service at a specified peak, the commodity charge as a charge for the quantity actually used. A flat-rate tariff can be viewed as a two-part rate in which the commodity charge is zero. Pareto-optimum pricing would seem to require that all parts of the tariff reflect the marginal cost of the appropriate product. If commodity costs are non-zero, flat rates are a *prima facie* departure from efficient pricing, in the absence of metering costs.[1] In this paper we examine the extent to which such departures may be justified by the existence of metering costs; the industry chosen for analysis is local telephone service. Two explanations for the frequently observed flat-rate structure are possible:

(1) That metering is undesirable because the costs of metering outweigh the gains in consumers' surplus which would result from bringing prices closer to marginal cost. As rationalized by one utility, "The package philosophy has developed because suppliers have found that they can satisfy the desires of consumers more economically in this way rather than by attempting to meet the exact needs of each consumer."[2]

(2) That the pricing policy is an example of the second type of behavior suggested by Averch and Johnson, in which the price of one output is set deliberately below cost in order to generate demand to support additional investment increasing overall permitted earnings under a rate base-rate of return regulatory constraint.[3]

Define

Q_c	as the number of local calls placed by consumers in an area, during a month,
Q_p	as the number of subscribers in the area,
P_c, P_p	as the respective prices per call and the monthly charge per phone,
C_c	as the (constant) marginal (overall) cost of making a call,
C_p	as the (constant) marginal (average) monthly cost of providing the connection of a single phone with the system, and
m	as the constant cost of measuring the duration and distance of a call.

We assume that the rest of the economy is in perfectly competitive equilibrium. Distribution effects are ignored.[4]

With metering, a Pareto-efficient price structure requires[5]

$$P_c = C_p + m$$

and

$$P_p = C_p .$$

---

[1] The problem of decreasing costs is recognized but ignored since the telephone industry, with which we are concerned, appears to have constant or increasing costs in the relevant directions.

[2] See [3], p. 12.

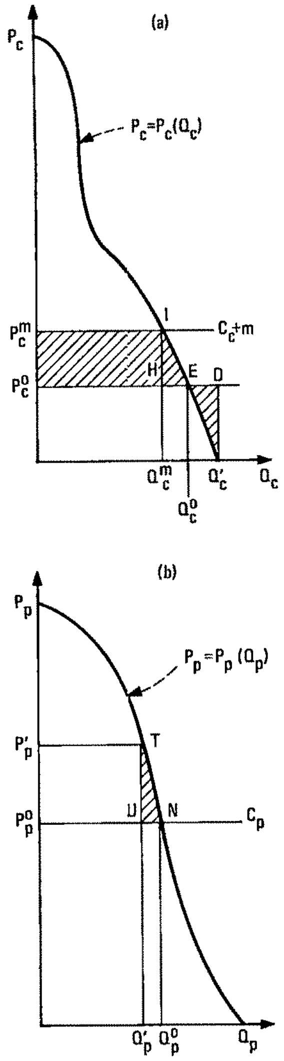
[3] See Averch and Johnson [1]. We are not concerned here with the other type of behavior postulated by Averch and Johnson, i.e., the choice of excessively capital-intensive technique in response to the constraint. See also Baumol and Klevorick [2].

[4] Redistribution effects favor heavy local users, though the extent of the resulting discrimination is at least in part offset by differential charges for party line, single residential, and business phones within the flat-rate structure typically employed.

[5] Problems of peak-load pricing, which would lead to a multiplicity of C_c and P_c values, are ignored in the interest of simplicity.

FIGURE 1

CONTINUOUS DEMAND FUNCTIONS



Instead, we observe

$$P_c = 0,$$
$$P_p > C_p,$$

with P_p increasing in real terms through time.[6]

Assume independent, continuous demand functions,[7] which appear in Figure 1:

$$P_c = P_c(Q_c) \tag{1}$$

$$P_p = P_p(Q_p)\,. \tag{2}$$

With no metering costs and marginal cost pricing, the number of calls would be Q_c^0 at a price P_c^0, the number of phones Q_p^0 at a price P_p^0. With metering and marginal cost pricing, the number of calls is Q_c^m at price P_c^m, while the number of phones remains Q_p^0 at price P_p^0. With a zero price on local calls, the number of calls rises to Q_c', while the number of phones is Q_p' at price P_p'.

Define B^m as consumers' surplus with metering and marginal cost pricing:

$$B^m = \int_0^{Q_p^0} P_p(\bar{Q}_p)d\bar{Q}_p - Q_p^0 \cdot C_p + \int_0^{Q_c^m} P_c(\bar{Q}_c)d\bar{Q}_c - Q_c^m(C_c + m)\,.$$

B' is consumers' surplus with a zero price on local calls and a price on phones in excess of marginal cost:

$$B' = \int_0^{Q_p'} P_p(\bar{Q}_p)d\bar{Q}_p - Q_p' \cdot C_p + \int_0^{Q_c'} P_c(\bar{Q}_c)d\bar{Q}_c - Q_c' \cdot C_c\,.$$

Define

$$\Delta s \equiv B^m - B'\,.$$

Δs is the change in consumers' surplus resulting from a move to a metered marginal cost pricing regime from an unmetered flat-rate one. Δs may be calculated as follows (areas described in parentheses refer to Figure 1):

$$\begin{aligned}
\Delta s \equiv {} & \int_{Q_p'}^{Q_p^0} P_p(\bar{Q}_p)d\bar{Q}_p - (Q_p^0 - Q_p') \cdot C_p \\
& - \int_{Q_c^m}^{Q_c'} P_c(\bar{Q}_c)d\bar{Q}_c + (Q_c' - Q_c^m) \cdot C_c - Q_c^m \cdot m \\
\equiv {} & \left[\int_{Q_p'}^{Q_p^0} P_p(\bar{Q}_p)d\bar{Q}_p - (Q_p^0 - Q_p') \cdot C_p\right] && \text{(area } TUN\text{)} \\
& + \left[(Q_c' - Q_c^0) \cdot C_c - \int_{Q_c^0}^{Q_c'} P_c(\bar{Q}_c)d\bar{Q}_c\right] && \text{(area } Q_c'DE\text{)} \\
& - \left[\int_{Q_c^m}^{Q_c^0} P_c(\bar{Q}_c)d\bar{Q}_c + (Q_c^0 - Q_c^m) \cdot C_c\right] && \text{(area } HIE\text{)} \\
& - Q_c^m \cdot m && \text{(area } P_c^0 P_c^m IH\text{)}.
\end{aligned}$$

---

[6] In Canada, over 1952–1967, the average annual increase, deflated using the Consumer Price Index, has been approximately 6.5 percent.

[7] Independence is assumed to permit calculation of consumers' surplus terms.

Metering increases welfare only if Δs is positive; if Δs is negative metering should be rejected. Evaluating Δs requires a great deal of information. For a simplified analysis assume that area $Q_c'DE =$ area HIE or

$$- \int_{Q_c^m}^{Q_c'} P_c(\bar{Q}_c)d\bar{Q}_c + (Q_c' - Q_c^m)\cdot C_c = 0 .$$

This would be the case, for example, if the demand curve for calls were linear and if m were equal to C_c. With this simplification, metering is desirable if the gain in consumers' surplus from additional phones with lower rates (area TUN) more than offsets the total cost of metering (area $P_c^0P_c^mIH$), i.e., if

$$\int_{Q_p'}^{Q_p^0} P_p(\bar{Q}_p)d\bar{Q}_p - (Q_p^0 - Q_p')\cdot C_p - Q_c^m\cdot m > 0 .$$

Under these assumptions we no longer need to estimate the demand function $P_c(Q_c)$ nor the marginal cost C_c.

Assume a constant elasticity demand function of the type

$$P_p = AQ_p^\sigma$$

where

$$\sigma < 0, \quad A > 0 \text{ and constant.} \tag{3}$$

Under our previous assumptions,

$$\Delta s = A\left(\frac{(Q_p^0)^{\sigma+1} - (Q_p')^{\sigma+1}}{\sigma + 1}\right) - (Q_p^0 - Q_p')\cdot C_p - Q_c^m\cdot m .$$

The metering rule becomes

$$\text{Meter if } m \leq \left[A\left(\frac{(Q_p^0)^{\sigma+1} - (Q_p')^{\sigma+1}}{\sigma + 1}\right) - (Q_p^0 - Q_p')\cdot C_p\right] \Big/ Q_c^m$$

$$\equiv R . \tag{4}$$

Implications of this rule have been examined under a number of demand conditions. Assume we have an unmetered local calling area with 500,000 telephones, on which an average of 30 calls per phone per month are made. Estimated average cost per month per phone is \$5.44. There are three possible metering systems of varying sophistication, which have increasing fixed cost as frequency, duration, and distance of calls are measured. Variable cost per call is constant and between \$0.01 and \$0.02 per call. R, the right-hand side of expression (4), has been calculated for a variety of assumed demand elasticities and unmetered monthly subscriber rates. Results appear in Table 1.

There is a case for metering only if R exceeds metering cost, assumed to be between \$0.01 and \$0.02 per call. The values shown in Table 1 suggest that these conditions are most likely to be met if demand elasticity is high or if the monthly charge required to cover calling costs is large relative to the cost of maintaining a phone in service.

These results are to some degree speculative, since we have not been able to obtain data which would permit an accurate estimate of the parameters involved. The values used are "guesstimates" provided

TABLE 1

CALCULATED RETURN PER CALL UNDER VARYING ELASTICITY CONDITIONS

CASE	DEMAND ELASTICITY FOR TELEPHONES	P'_p	R
1	−0.25	$7.00	$0.0016
2	−0.50	7.00	0.0030
3	−1.50	7.00	0.0044
4	−2.00	7.00	0.0140
5	−0.25	8.00	0.0040
6	−0.50	8.00	0.0082
7	−1.50	8.00	0.0281
8	−2.00	8.00	0.0749
9	−0.25	9.00	0.0071
10	−0.50	9.00	0.0144
11	−1.50	9.00	0.0531
12	−2.00	9.00	0.7273

Q'_p = 500,000 TELEPHONES PER MONTH

C_p = $5.44 PER MONTH

Q_c^m = 30 CALLS PER PHONE PER MONTH

by the engineering department of a major Canadian telephone utility. Experimentation with changing values suggests that increasing Q_c reduces R if no change in C_p is necessary. The assumption that $C_c = m$ is only a convenience; if $C_c < m$, the conditions under which metering becomes preferable are more stringent in terms of demand elasticities and/or P_p' than those shown in the table. With Averch-Johnson (A-J) behavior, we would expect to find flat rates universally used. The observed pattern of choices with respect to rate structures by U. S. telephone utilities thus appears more consistent with rational pricing in the face of metering costs than with A-J behavior.

Casual observation suggests that metering of calls is much more frequent in European and other overseas systems than it is in North America. Again, this is not inconsistent with the likelihood that demand elasticities for telephone service in these markets are greater than they are in North America.

References

1. Averich, H. and Johnson, L. L. "Behavior of the Firm Under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–69.
2. Baumol, W. J. and Klevorick, A. K. "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–90.
3. Trans-Canada Telephone System. *Canada's Telephone Industry in Perspective.*

A state preference approach to public utility pricing and investment under risk

Stephen C. Littlechild

Graduate Centre for Management Studies
University of Birmingham

This note formulates the problem of public utility pricing and investment under risk in terms of the "state preference" approach. Price is assumed to be set after, rather than before, demand is observed. The model is more general, and the results are obtained more immediately, than in a previous formulation.

1. Introduction

■ Brown and Johnson analyzed the problem of public utility pricing and output under risk.[1] They obtained a characterization of optimal pricing and investment policy, and a comparison between the resulting prices and outputs and those which would obtain in a world of certainty.

This note provides an analysis of the problem in terms of the so-called "state preference" approach. This approach has three major advantages. First, the results are obtained more easily. Second, the analysis does not depend on additive or multiplicative relationships among the stochastic demand curves, nor upon linearity of these curves. Third, two immediate generalizations of the model are available, namely, to the case of interdependent demand curves and to the case where supply as well as demand is stochastic.

There is one significant difference between the assumptions of this and the previous model. Brown and Johnson assumed that price was to be set before demand was observed. Here we assume it is set afterwards. We discuss first the reason for the difference in assumptions. In Section 3 we present the state preference model and derive the characterization of optimal policy. Section 4 contains some observations on the resulting prices, and Section 5 makes the comparison with the case of non-stochastic demand. In Section 6 the two generalizations are indicated.

2. Assumptions

■ In their reply to two critics, Brown and Johnson wrote:

> We assumed that the presence of uncertainty requires the public utility to set both price and capacity output before the position of the demand function is known. . . . This assumption alone is responsible for the conclusion that optimal price should equal short-run marginal cost.[2]

This is not entirely correct. The crucial assumption is that

> when output falls short of demand, available output is distributed to individual customers on the basis of their valuation of the marginal unit of the good.[3]

---

[1] In [1].

[2] [2], p. 489.

[3] [1], p. 123.

Salkever[4] pointed out that if the second assumption is made, then any price between marginal operating cost and the market-clearing price, where the latter is above the former, is as (allocatively) efficient as any other price between those two limits. If the first assumption is also made, then price must be set equal to the lower limit, because the upper limit is as yet unknown. However, if the second assumption is not made, then, as Turvey[5] recognized, the disadvantage of setting too high a price, namely the opportunity cost of holding idle capacity in times of low demand, would have to be balanced against the disadvantage of setting too low a price, namely the opportunity cost of inefficiently distributing scarce output in times of high demand.

Evidently, if price is set before demand is observed, then some assumption must be made about the rationing procedure in the event that demand exceeds capacity. The assumption that Brown and Johnson make is the simplest to handle analytically, but from the point of view of resource allocation it is equivalent to assuming that price is set after demand is observed. The latter assumption is even more tractable analytically, which is why it is used herein.[6]

Moreover, contra Johnson and Brown, there do not seem to be practical grounds for rejecting the case where price is set after demand is observed. Gas supplies are already offered on an interruptible basis at substantial price reduction, so that effectively for such consumers, "on the coldest night of the year the price of gas . . . [is] raised to an all time high!"[7] Vickrey[8] has proposed a real-time congestion-dependent pricing scheme for telephones. Stevens and Shutler argue that "the lag between orders and deliveries allows steel producers to change prices in accordance with the likely changes in the future use of their steel capacity, as indicated by their order books."[9] Airline tickets sold on standby basis constitute another simple example.

This is not of course to argue that all prices can be set after demand is observed; many cannot.[10] A fruitful topic for future research would seem to be a theoretical and empirical study of the effects of the various means of non-price rationing hitherto adopted by public utilities.

3. The model

■ Proceeding to develop the model, let subscripts $s = 1, \ldots, n$ denote the finite set of possible states of the world, with probabilities θ_s of occurring, where $\theta_s \geq 0$ and $\sum_s \theta_s = 1$. For each state s, let x_s denote output and $f_s(x_s)$ its marginal valuation function or (inverse) demand curve. These functions are assumed downward sloping, i.e.,

$$f_s'(x_s) < 0, \tag{1}$$

---

[4] See [7].

[5] See [9].

[6] The more realistic case of random rationing is treated by Marchand in his analysis of telephone pricing [5].

[7] Brown and Johnson [2], p. 490.

[8] In [10].

[9] [8], p. 5.

[10] Moreover, as Mr. John Kay has observed to me, the demand for many products becomes increasingly inelastic as the planned time of consumption approaches.

but not necessarily linear. Additive or multiplicative relationships between these functions would represent the special cases treated by Brown and Johnson. Capacity is denoted y. Marginal operating and capacity costs are assumed constant and represented by b and β, respectively. An asterisk will denote the optimal value of a variable.

The problem is to choose output x_s in each state s and capacity y so as to maximize the expected value of consumers' evaluation of output net of operating costs, less the cost of capacity

$$\Sigma_s \theta_s \left\{ \int_0^{x_s} f_s(\xi_s) d\xi_s - b x_s \right\} - \beta y \tag{2}$$

subject to the capacity constraint in each state

$$x_s \leq y, \quad s = 1, \ldots, n. \tag{3}$$

Evidently capacity is supplied jointly to all states, so that the problem is formally the same as other problems (e.g., peak-load pricing, depreciation, etc.) involving joint costs.[11]

Let v_s denote the dual variables to these constraints and in what follows transform by setting $\theta_s u_s = v_s$ and dividing where appropriate by θ_s. Assuming strictly positive optimal investment y^* and output x_s^* in all states, the following Kuhn-Tucker conditions hold:

$$f_s(x_s^*) = b + u_s^*, \quad s = 1, \ldots, n \tag{4}$$

$$\sum_s \theta_s u_s^* = \beta, \tag{5}$$

where

$$u_s^* = 0 \quad \text{if} \quad x_s^* < y^*, \quad s = 1, \ldots, n. \tag{6}$$

Optimal pricing and investment policy may be characterized as follows. Price p_s^* in state s is set equal to $f_s(x_s^*)$. If demand at price equal to marginal operating cost would have exceeded capacity, then capacity output is produced ($x_s^* = y^*$) and $p_s^* = f_s(x_s^*) = f_s(y^*)$ is the market-clearing price in state s. By (4), u_s^*, which is the difference between market-clearing price and marginal operating cost, may be interpreted as the marginal opportunity cost (or value) of capacity in state s. In states where demand at price equal to marginal operating cost would not exceed capacity, price is set equal to marginal operating cost ($p_s = b$), and in such states marginal opportunity cost is zero, by (6). The optimal investment criterion (5), is that capacity is purchased to the point where its expected marginal (opportunity) value equals its marginal cost.

4. Observations

■ Certain observations on the above pricing policy are in order. First, expected price equals marginal operating cost plus marginal capacity cost, for

$$\sum_s \theta_s f_s(x_s^*) = \sum_s \theta_s (b + u_s^*) = b + \beta \quad \text{by (4) and (5)}. \tag{7}$$

Moreover, expected total revenue equals total cost. Ignoring operating costs, which are paid for as incurred, we have

$$\sum_s \theta_s x_s^* u_s^* = y^* \sum_s \theta_s u_s^* = \beta y^* \quad \text{by (6) and (5)}. \tag{8}$$

---

[11] See Littlechild [4].

Both these results depend on constant marginal costs b and β. The expected price $(p = b + \beta)$ may be quoted to customers *before* demand is observed, provided that these customers are intra-marginal consumers in each state. Then the remaining output in each state may be sold at the previously determined market-clearing prices $b + u_s^*$.

Finally, the dual variables may be given an interesting interpretation in terms of the contingent future markets mentioned by Brown and Johnson. If state s occurs, then u_s^* is the marginal value of capacity. $v_s^* = \theta_s u_s^*$ therefore represents the expected value of a conditional claim to capacity in state s alone. Thus, the intra-marginal consumer of the previous paragraph is being sold the unconditional right to use the commodity at a price equal to marginal operating cost plus the expected value of conditional claims to capacity summed over all states. By contrast, an "interruptible" consumer charged on the same basis would be sold the right to use gas in all but the peak state, say $s = 1$, at a price of p' given by

$$p' = b + \frac{\sum_{s \neq 1} \theta_s u_s^*}{1 - \theta_1} < b + \beta .$$

5. Comparison with non-stochastic demand

■ In this section we shall derive and discuss the proposition of Brown and Johnson that optimal capacity is not less than that which would be installed if the demand function were non-stochastic. The "equivalent" non-stochastic (inverse) demand function $\bar{f}(x)$ is defined as the expected value of the stochastic (inverse) demand function

$$\bar{f}(x) = \sum_s \theta_s f_s(x) . \qquad (9)$$

With this non-stochastic demand function, optimal capacity $\bar{y}$ is characterized by

$$\bar{f}(\bar{y}) = b + \beta. \qquad (10)$$

A comparison of y^* and $\bar{y}$ is required. We have

$$\begin{aligned} \bar{f}(y^*) &\equiv \sum_s \theta_s f_s(y^*) && \text{by definition (9)} \\ &\leq \sum_s \theta_s f_s(x_s^*) && \text{by (3) and (1)} \\ &= b + \beta && \text{by (7)} \\ &= \bar{f}(\bar{y}) && \text{by (10)} . \end{aligned}$$

But $\bar{f}(\bar{y}^*) \leq \bar{f}(\bar{y})$ implies $y^* \geq \bar{y}$ by (1). That is, capacity with stochastic demand is not less than capacity with "equivalent" non-stochastic demand.

Under what conditions will these two capacities be equal?—evidently, if and only if for all states optimal output x_s^* equals capacity y^*. Conversely, capacity with stochastic demand is strictly greater if and only if there is a positive probability of a state in which demand at price equal to marginal operating cost is less than capacity, that is, if there is a true "off-peak" state.

An intuitive explanation for these results can be obtained by comparing the effective demands for capacities under the two assumptions. Effective demand for capacity in the non-stochastic case is given by $\bar{f}(y) - b$. In the stochastic case output can be curtailed in states of low demand, because price is set equal to marginal operating cost whenever this exceeds the market clearing price level. Effective demand for capacity is therefore given by $\sum_s \theta_s \max\{f_s(y) - b, 0\}$. But

$$\sum_s \theta_s \max\{f_s(y) - b, 0\} \geq \sum_s \theta_s f_s(y) - b \equiv \bar{f}(y) - b\,. \tag{11}$$

In other words, effective demand is as high or higher in the stochastic case. In a world in which capacity output has to be produced in each state, and where total consumption of this output is compelled, optimal capacities under risk and under certainty would be equal.

6. Two generalizations

■ Finally, two immediate extensions to the model suggest themselves.

Demand in any state is likely to depend not only upon price in that state but also upon (expected) prices in all other states, because the services of consumption equipment are joint to all states. This can be formally treated by defining inverse demand $f_s(x_1, \ldots, x_n)$ as a function of outputs in all states and maximizing the line integral

$$\int \sum_s f_s(\xi_1, \ldots, \xi_n) d\xi_s - b \sum_s x_s - \beta y \tag{2'}$$

from the point $(0, \ldots, 0)$ to the point $(x_1, \ldots, x_p)$ subject to constraints (3). This technique, proposed by Hotelling,[12] is discussed by Pressman.[13] The qualitative results of the present analysis are not substantially altered.

In certain industries supply as well as demand is stochastic. For example, supplies of water and of hydroelectric power are both likely to be lower in precisely those states of high temperature when demands for irrigation and for air-conditioning, respectively, are highest. Demand and supply need not be thus correlated, of course. These cases can be treated by defining α_s as the proportion of nominal capacity which is available in state s and rewriting the capacity constraints as

$$x_s \leq \alpha_s y, \quad s = 1, \ldots, n. \tag{3'}$$

The characterization of optimal capacity becomes

$$\sum_s \theta_s \alpha_s u_s^* = \beta\,. \tag{5'}$$

Assuming $0 \leq \alpha_s \leq 1$, each unit of investment produces on the average less available capacity than before. In general, nominal capacity will be lower and prices higher than before. The shadow prices are likely to be useful for evaluating inter-utility transfers of output which are increasingly taking place.

References

1. Brown, G., Jr. and Johnson, M. "Public Utility Pricing and Output Under Risk." *The American Economic Review*, Vol. 59, No. 1 (March 1969), pp. 119–28.

---

[12] In [3].

[13] See [6].

2. ———. "Public Utility Pricing and Output Under Risk: Reply." *The American Economic Review*, Vol. 60, No. 3 (June 1970), pp. 489–90.
3. HOTELLING, H. "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates." *Econometrica*, Vol. 6, No. 2 (July 1938), pp. 242–69.
4. LITTLECHILD, S. C. "Marginal-Cost Pricing with Joint Costs." *Economic Journal*, Vol. 80, No. 318 (June 1970), pp. 323–25.
5. MARCHAND, M. "Economic Principles and Telephone Rates." Working Paper No. 251, Center for Research in Management Science, University of California, Berkeley, April 1968.
6. PRESSMAN, I. "A Mathematical Formulation of the Peak-Load Pricing Problem." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autum 1970), pp. 304–26.
7. SALKEVER, D. S. "Public Utility Pricing and Output Under Risk: Comment." *The American Economic Review*, Vol. 60, No. 3 (June 1970), pp. 487–88.
8. STEVENS, J. N. and SHUTLER, M. F. "Marginal Cost Pricing in the U. K. Steel Industry." Paper presented in the 17th International Conference of The Institute of Management Science, London, July 1, 1970.
9. TURVEY, R. "Public Utility Pricing and Output Under Risk: Comment." *The American Economic Review*, Vol. 60, No. 3 (June 1970), pp. 485–86.
10. VICKREY, W. "Responsive Pricing of Public Utility Services." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 337–46.

User charges as a means for pollution control: the case of sewer surcharges

Don Ethridge

Assistant Professor of Economics
University of Missouri at Rolla

Some cities in the U. S. are using sewer surcharges based on the strength of wastes discharged to induce industrial firms to decrease their discharges of water-carried wastes. Many other cities are either in the process of instituting sewer surcharges or are considering instituting them. This study reports the estimates of the responses of the poultry and beet processing industries to sewer surcharges which were derived using two different estimation procedures. A surcharge of $0.02 per pound of biochemical oxygen demand is estimated to decrease the total pounds by 75 percent for beet processing firms and approximately 25 percent for poultry processing firms.

■ In recent years, the regulation of pollution of public waters has begun to be put into effect. This regulation has often taken the form of stream standards with the stream being classified according to its best alternative use, "best" being defined by a waterway agency or authority, and quality standards established that will render the waterway suitable for such purposes. As a result, industrial plants and municipalities located along waterways are not allowed to discharge wastewater into the waterways such as to cause the quality of the water to fall below the defined level.

In many cities, the problem of meeting such stream standards is complicated by industrial firms' use of the municipal sewer system, since some of these firms discharge highly concentrated wastes. Industrial firms pay a charge for sewer services, but in most cases the cost of sewer service to the firm is a constant proportion of the cost of water service; the service charge is dependent upon the quantity of water used instead of the quantity of wastes discharged. Consequently, there is no economic incentive in the sewer service rate structure for firms to reduce their waste discharges, resulting in an overload of cities' treatment facilities. This has led some cities to impose sewer surcharges, i.e., place a price on waste discharges, on industrial firms that discharge strong wastes in an attempt to (1) distribute sewage treatment costs more equitably among residential and industrial users and (2) reduce the volume of wastes being treated by municipal plants or to (3) help finance the construction and operation of new municipal treatment facilities.[1]

This paper is part of the author's Ph.D. dissertation at North Carolina State University. Helpful comments were received from J. A. Seagraves, North Carolina State University, and John Rafferty, National Center for Health Services Research.

[1] See P. T. Hickman, "The Introduction of an Industrial Waste Surcharge Program," paper presented at the annual meeting of the Missouri Water Pollution Control Association, St. Louis, Missouri, 1964.

While many cities across the country are considering instituting sewer surcharges, little is known of industry's response. If firms reduce their wastes, city sewage treatment systems may not need to expand, or at least not as rapidly. Also, if firms reduce their wastes when faced with surcharges, it would imply that firms not using municipal sewer services would respond in the same manner if they were made to pay for wastes discharged. The purpose of this paper is to present the estimated responses of two industries to sewer surcharges.

1. Definitions and conceptual framework

■ One of the basic purposes of sewer surcharges is to make the firm responsible for its own strong wastes. The definition of strong wastes for present purposes is any quantity of wastes of the sewage greater than some predetermined amount (defined in terms of some common measure for the different waste elements) per unit of water. The firm has four alternatives available: (1) to reduce the strength of its wastewater to the specified strength by measures such as in-plant treatment, production process changes, etc., (2) to pay a surcharge to the city to clean its wastewater to the specified strength, (3) to dilute its wastewater to the specified strength with clean water, or (4) any combination of (1), (2), and (3). The type of surcharge structure dealt with here is that which is most predominant, and it is assumed that the surcharge is levied on the basis of BOD (biochemical oxygen demand) alone.[2] The rules facing the firm allow it to discharge a given number of pounds of BOD as normal wastes, defined in terms of ppm (parts per million) of BOD, on which no surcharge is levied; the charge for normal wastes is in the regular sewer charge.

Consider a firm which is using a municipal water and sewer system and assume a given non-waste-induced technology,[3] a given plant and volume of production, and given input prices. All of the firm's water-carried wastes are either (1) normal wastes, (2) waste removed by the city (with the surcharge paid on it), or (3) waste removed by the firm by in-plant waste removal before it enters the sewer system. Thus, for the firm,

$$\mathrm{WCW} \equiv \mathrm{EMT} + \mathrm{PR} + \mathrm{NW}(\mathrm{W}) \equiv \mathrm{TMT} + \mathrm{PR}, \qquad (1)$$

---

[2] BOD indicates the oxygen demand which organic wastes place on a stream or waterway; a large BOD load depletes a stream of oxygen and affects aquatic organisms which require oxygen to live. It is a measure which is used to convert many different kinds of effluents to a common unit of measurement. Although many cities levy surcharges on BOD and other measures of wastes (such as suspended solids), the theory is developed using BOD alone as the measure of wastes in order to simplify matters. A given quantity of effluent may contain suspended solids which also constitute BOD. However, little can be said in general about the relative amounts of each. If the relationships were fixed and known, all wastes could be converted to one measure.

[3] The term "non-waste-induced technology" is used to distinguish those technological factors which are made explicitly to reduce the amount of BOD entering the sewer from those made for other reasons but which may nevertheless affect the amount of water-carried wastes. Those innovations made to reduce the pounds of BOD entering the sewer are part of those actions that constitute in-plant waste reduction; they would not be made in the absence of the surcharge. Non-waste-induced technology refers to those technical innovations made to save labor, time, etc., but which may also affect water-carried wastes.

where

WCW = water-carried wastes measured in pounds of BOD per unit of time; this refers to those wastes that would be in the wastewater in the absence of a surcharge;

EMT = excess municipal treatment or the pounds of BOD per unit of time over and above the pounds of BOD allowed as normal wastes; the surcharge applies only to EMT;

PR = in-plant waste removal measured in pounds of BOD per unit of time; this refers to those wastes that are removed by the firm because of the surcharge;

NW (W) = normal wastes measured in pounds of BOD per unit of time and which is covered by the standard sewer rate structure; NW is in ppm of BOD and W is in million pounds of water; and

TMT = EMT + NW (W) = total municipal treatment done by the city for the firm measured in pounds of BOD per unit of time.

Assume that the firm's goal is to maximize its net revenue under the surcharge. If the firm does not change its output of product,[4] this is equivalent to minimizing its cost of waste reduction, which can be described by the condition that the marginal cost of in-plant reduction equals the surcharge; $MC_{pr} = P_{mt}$, where MC_{pr} is the net marginal cost of BOD removal by the firm and P_{mt} is the surcharge (both are stated in cents per pound of BOD). This MC_{pr} is the marginal cost of removing BOD from the wastewater after adjustments in costs have been made for variations in productive inputs, including water, due to the in-plant reduction and represents the least-cost condition for in-plant BOD removal. The firm's marginal cost of in-plant waste reduction is expected to increase with the amount of in-plant reduction; $MC_{pr} = f(\text{PR})$, $\frac{dMC_{pr}}{d\text{PR}} > 0$.[5] It is assumed that the form of the relationship can be represented as

$$MC_{pr} = \beta_0 e^{\beta_1 PR} = P_{mt}, \tag{2}$$

where $e = 2.71828$ and β_0 and β_1 are parameters. The amount of water used by the firm, i.e, the firm's demand for water, may also be affected by the surcharge (the firm may use waste dilution);

---

[4] While the firm's optimum response to the surcharge may be to decrease its output of product to some degree, an offsetting factor may be the recovery of by-products from the production process. The assumption of constant product output is made for simplicity. If the firm does reduce its production, the effect would be to make the firm's demand for municipal treatment more elastic.

[5] In the process of reducing waste discharges, the firm is expected to engage in those practices which are the least-cost practices first, but as greater levels of waste reduction are performed, the costs become progressively greater, i.e., each additional unit of waste is more difficult and more costly to remove. This is the type of relationship found by G. O. G. Löf and A. V. Kneese, in *The Economics of Water Utilization in the Beet Sugar Industry*, Resources for the Future, Inc. (Baltimore, Md.: The Johns Hopkins Press, 1968).

W $= f(P_{mt})$. Thus, solving for PR from (2) and substituting into (1),

$$\mathrm{EMT} = \mathrm{WCW} - \frac{1}{\beta_1}\ln\left(\frac{1}{\beta_0}P_{mt}\right) - \mathrm{NW}\,f(P_{mt})\,, \qquad (3)$$

$$\mathrm{TMT} = \mathrm{WCW} - \frac{1}{\beta_1}\ln\left(\frac{1}{\beta_0}P_{mt}\right), \qquad (4)$$

where ln indicates the logarithm to the base e. Equations (3) and (4) specify the firm's demand for EMT and TMT respectively. From these, the elasticities of demand with respect to the sewer surcharge are

$$^{\eta}\mathrm{EMT}/P_{mt} = -\frac{1}{\beta_1\mathrm{EMT}} - \frac{\mathrm{NW}P_{mt}}{\mathrm{EMT}}\frac{d\mathrm{W}}{dP_{mt}}\,, \qquad (5)$$

$$^{\eta}\mathrm{TMT}/P_{mt} = -\frac{1}{\beta_1\mathrm{TMT}}\,. \qquad (6)$$

The relevance of these elasticities lies in their implications for the enforcement of surcharges by city governments. While the elasticity of TMT with respect to the surcharge describes changes in the industrial waste load on a municipal treatment plant, the elasticity of EMT is useful for summarizing the effects on total revenue from the surcharge. If the absolute value of $^{\eta}\mathrm{EMT}/P_{mt}$ is less than one, then an increase in P_{mt} would result in an increase in the total surcharge paid by the firm. Estimates of $^{\eta}\mathrm{TMT}/P_{mt}$ would provide relevant information on the expected decrease in TMT from an increase in the surcharge. Note that it is possible for TMT to be inelastic while EMT is elastic. However, the city may alter this situation by redefining NW at some lower waste strength, lowering $^{\eta}\mathrm{EMT}/P_{mt}$; if NW is defined as zero ppm of BOD, then (5) and (6) become the same. Cities using sewer surcharges therefore have two additional policy variables with which to affect firms' use of municipal treatment: the surcharge and the definition of normal wastes.

2. Estimates of industry responses to surcharges

■ Two approaches to the estimation of the industrial response to sewer surcharges were considered. In the "synthetic" approach, *a priori* knowledge was used to estimate the response of the beet processing industry. In the second approach a regression model for the poultry processing industry was formulated and the response estimated from sample data.

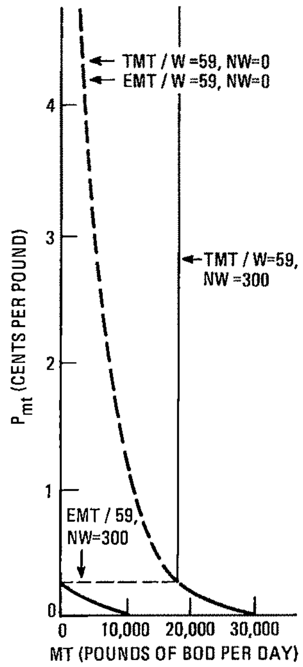
☐ **Synthetic derivation of the demand for municipal treatment: beet sugar industry.** The data for this example come from an analysis of beet sugar processing by Löf and Kneese,[6] who presented an incremental cost curve for a typical beet sugar processing plant as a function of the pounds of BOD removed per day. Their typical plant is one which processes 2,700 tons of beets per day, uses approximately 59 million pounds of water per day, and generates a total of 31,600 pounds of BOD per day.

---

[6] Lof and Kneese, *op. cit.*

FIGURE 1

DEMAND FOR MUNICIPAL WASTE TREATMENT FOR A 2,700 TON PER DAY BEET SUGAR PROCESSING PLANT



WATER USE IS CONSTANT AT 59 MILLION POUNDS PER DAY.
WCW = 31,600 POUNDS OF BOD PER DAY
NW = 300 ppm OF BOD

The following least-squares estimate represents their incremental cost curve:

$$MC_{pr} = 0.01965e^{(0.000194)PR}. \tag{7}$$

Assuming that NW = 300 ppm of BOD and that water intake does not change, then substituting into equations (3) and (4),

$$\mathrm{EMT} = 13{,}900 - \frac{1}{0.000194}\ln\left(\frac{1}{0.01965}P_{mt}\right), \tag{8}$$

$$\mathrm{TMT} = 31{,}600 - \frac{1}{0.000194}\ln\left(\frac{1}{0.01965}P_{mt}\right). \tag{9}$$

These demand curves are shown in Figure 1. The EMT curve intercepts the vertical axis and becomes perfectly inelastic at a surcharge of $0.29 per pound of BOD. This is much lower than most municipal surcharges, but for purposes of illustration, if P_{mt} is, say, $0.25 per pound of BOD, EMT = 791 and TMT = 18,491. Therefore, for these conditions, the elasticities would be $^{\eta}\mathrm{EMT}/P_{mt} = -6.517$ and $^{\eta}\mathrm{TMT}/P_{mt} = -0.279$. Note that TMT becomes vertical at the same P_{mt} where EMT = 0; when the firm no longer has to pay a surcharge on its waste load because of the defined NW, there is no incentive to reduce the pounds of BOD discharged from additional increases in P_{mt}.

☐ **Regression estimate of the demand for municipal treatment: poultry processing industry.** In formulating a regression model with which to estimate the response to surcharges for poultry firms, the same theoretical model applies as for the synthetic estimation procedure, but several additional factors require consideration. Adjustments were made: (1) to convert to municipal treatment per 1,000 birds processed, (2) to adjust for differences in non-waste-induced technology over time and across firms, and (3) to adjust for differences in the types of surcharge structures across firms; some cities levy surcharges on BOD only and some cities on both BOD and suspended solids.[7]

From the identity

$$\frac{\mathrm{EMT}}{Y} \equiv \frac{\mathrm{WCW}}{Y} - \frac{\mathrm{PR}}{Y} - \mathrm{NW}\left(\frac{\mathrm{W}}{Y}\right) - \mathrm{NW}\left(\frac{\mathrm{W}}{Y}\right), \tag{10}$$

where

$\frac{\mathrm{EMT}}{Y}$ = excess pounds of BOD per 1,000 birds processed,

$\frac{\mathrm{TMT}}{Y}$ = total pounds of BOD per 1,000 birds,

---

[7] Some part of suspended solids also constitutes a BOD load. A problem exists, however, in that a conversion factor for converting suspended solids to its BOD equivalent may not exist and if it does exist is not known. It is assumed that two pounds of suspended solids constitute one pound of BOD, i.e., $2\,\mathrm{SS} = \mathrm{BOD}$. Converting from a surcharge on both BOD and suspended solids to the equivalent charge on BOD alone,

$$\begin{aligned} \text{Total surcharge} &= P_{mt}(\mathrm{BOD}) + P_{ss}(\mathrm{SS}) \\ &= P_{mt}(\mathrm{BOD}) + P_{ss}(\tfrac{1}{2}\mathrm{BOD}) \\ \text{Surcharge per pound of BOD} &= \frac{\text{Total surcharge}}{\mathrm{BOD}} = P_{mt} + \tfrac{1}{2}P_{ss} = P_{mt}^{*}. \end{aligned}$$

$\frac{\text{WCW}}{Y}$ = pounds of water-carried BOD per 1,000 birds processed that would exist in the absence of a surcharge,

$\frac{\text{PR}}{Y}$ = pounds of BOD removed by the firm per 1,000 birds processed,

$\text{NW}\left(\frac{\text{W}}{Y}\right)$ = pounds of BOD classified as normal per 1,000 birds processed, and

$\frac{\text{W}}{Y}$ = million pounds of water per 1,000 birds processed.

The structural equations used were:

$$\frac{\text{WCW}}{Y} = \beta_0 + \beta_2 T_2 + \beta_3 MC_{w+s}\,, \tag{11}$$

$$\frac{\text{PR}}{Y} = \frac{1}{\beta_4} \ln \frac{1}{\beta_{14}} (P_{mt}{}^*)\,, \tag{12}$$

$$\frac{\text{W}}{Y} = \beta_7\, e^{\beta_8 P_{mt}{}^* + \beta_9 MC_{w+s} + \beta_{10} T_2}\,, \tag{13}$$

$$\frac{\text{TMT}}{Y} = \beta_0 + \frac{1}{\beta_4} \ln\beta_{14} + \beta_2 T_2 + \beta_t MC_{w+s} - \frac{1}{\beta_4} \ln(P_{mt}{}^*)\,, \tag{14}$$

$$\frac{\text{EMT}}{Y} = \beta_0 + \frac{1}{\beta_4} \ln\left(\frac{1}{\beta_{14}}\right) + \beta_2 T_2 + \beta_3 MC_{w+s} - \frac{1}{\beta_4} \times \ln(P_{mt}{}^*) - \text{NW}\beta_7 e^{\beta_8 P_{mt}{}^* + \beta_9 MC_{w+s} + \beta_{10} T_2}\,, \tag{15}$$

where

T_2 = automatic gizzard splitting (1 denotes presence, 0 denotes absence),

MC_{w+s} = the marginal cost of water and normal sewer services measured in cents per million pounds of water,

P_{mt} = the sewer surcharge in cents per pound of BOD, and

P_{ss} = the sewer surcharge in cents per pound of suspended solids.

Data used for the regression model were observations on plant output, technology employed in the plant, the surcharge (on BOD and on suspended solids), the quantity of water used by the plant, and pounds of BOD entering the sewer, all across firms and over time. The sample was limited to poultry processing firms located in cities that had been providing for the enforcement of municipal sewer surcharges for more than five years. Of the 23 plants in 17 cities and 12 states chosen as a sample, data were obtained from only five plants and a total of 27 observations from the five firms was obtained.

The following equations are presented as the most realistic estimates of $\frac{\text{TMT}}{Y}$ and $\frac{\text{EMT}}{Y}$ for the poultry processing industry from these data:

$$\frac{\text{TMT}}{Y} = \underset{}{68.86} + \underset{(9.89)^*}{13.41 T_2} - \underset{(0.0034)^*}{0.0077 MC_{w+s}} - \underset{(6.57)^{**}}{16.25 \ln(P_{mt}{}^*)} \qquad R^2 = 0.50 \tag{16}$$

$$\ln\frac{\text{W}}{Y} = -\ 0.16933 - 0.1987P_{mt}{}^* - 0.0002MC_{w+s} + 0.3777T_2 \quad (17)$$

$$\qquad\qquad (0.0423)^{***} \quad (0.0001)^{**} \quad (0.1359)^{***} \quad R^2{=}0.55$$

Standard errors are shown in parentheses below the estimated coefficients. *, **, and *** denote significance at the 0.10, 0.05, and 0.01 levels, respectively. Assuming that $MC_{w+s} = 3144.3$ (the mean value), $T_2 = 0$, and $P_{mt}{}^* = 2.194$ (the mean):

$$\frac{\eta\text{TMT}}{Y}\Big/P_{mt}{}^* = -\ 0.510$$

$$\frac{\eta\text{W}}{Y}\Big/MC_{w+s} = -\ 0.629$$

$$\frac{\eta\text{W}}{Y}\Big/P_{mt}{}^* = -\ 0.436$$

$$\frac{\eta\text{EMT}}{Y}\Big/P_{mt}{}^* = -\ 0.618$$

$$\frac{\eta\text{EMT}}{Y}\Big/MC_{w+s} = -\ 0.952$$

$$\frac{\eta\text{EMT}}{Y}\Big/\text{NW} = -\ 1.478$$

$$\frac{\eta\text{TMT}}{Y}\Big/MC_{w+s} = -\ 0.759.$$

From the foregoing analysis, the following conclusions were drawn with respect to the poultry processing industry:

(1) The surcharge on BOD does significantly affect the total pounds of BOD treated by the city; the elasticity of pounds of BOD discharged per 1,000 birds with respect to the surcharge on BOD is estimated to be −0.5 at the mean surcharge.

(2) The surcharge on BOD probably decreases the quantity of water used per 1,000 birds. However, the size of the effect is uncertain since the estimates appear to be sensitive to assumptions made about the proportion of the suspended solids surcharge applicable as a BOD surcharge. The elasticity estimate obtained indicates that a one-percent increase in the surcharge would result in a 0.4-percent decrease in the water used per 1,000 birds.

(3) The marginal cost of water-related services does significantly affect the total pounds of BOD per 1,000 birds processed entering the wastewater. It appears that a $0.01 increase in the marginal cost of water and sewer services in cents per million pounds of water results in a decrease of about 0.01 pounds of BOD per 1,000 birds.

(4) There is a significant relationship between the marginal cost of water-related services and the quantity of water used per 1,000 birds within the range of observed values. The elasticity of water used per 1,000 birds with respect to the price of water is about −0.6.

(5) An increase in the surcharge would increase the total surcharge bill and reduce the water bill of poultry processing firms. $\frac{\text{EMT}}{Y}$ is inelastic with respect to the surcharge but more elastic than $\frac{\text{TMT}}{Y}$. $\frac{\text{EMT}}{Y}$ is more inelastic because of the negative effect of P_{mt} on $\frac{\text{W}}{Y}$, *ceteris paribus.*

(6) The presence of automatic gizzard splitting seems to affect the total pounds of BOD per 1,000 birds processed entering the wastewater and the quantity of water used. The presence of this operation appears to increase the waste load by about 13 pounds of BOD per 1,000 birds processed.

3. Summary and conclusions

■ The elasticity of demand for total municipal treatment in the beet sugar processing industry is approximately −0.7 at the prevailing surcharge (about $0.02 per pound of BOD) if normal wastes are defined such that the surcharge would actually apply. The corresponding elasticity of demand for total municipal treatment in the poultry processing industry is approximately −0.5 at the prevailing surcharge. Surcharges appear to induce poultry processing firms to use less water, and the number of pounds of wastes discharged by poultry processing firms also appears to be responsive to the cost of water; the elasticity of water used per 1,000 birds with respect to the surcharge was −0.4, indicating that water use for the firms sampled decreased as the surcharge increased. In addition, it appears that poultry processing firms' wastes are more responsive to the cost of water and the definition of normal wastes than to the surcharge.

For poultry processing, EMT was more elastic with respect to P_{mt} than was TMT; this occurs because of the net decrease in water use due to the surcharge. The introduction of a surcharge program may affect water use by the firm in two opposing ways: (1) The restriction of having to reduce its waste strength may induce firms to engage in waste dilution; and (2) the firm may decrease water use because of in-plant changes associated with PR. This study indicates that the net effect for poultry processing has been to decrease water use. Whether water use would decrease for beet processing is not known since the synthetic approach did not yield this information.

The preceding analysis assumes that the firm has the alternative of having the municipal system treat the wastes and paying a surcharge. In some cases, cities require firms to produce the waste reduction but do not provide the service of treating excess wastes. In this instance, the firm would have only two alternatives: in-plant reduction of wastes and waste dilution. However, this approach to the problem would encourage greater water use by firms,[8] a choice which may be particularly undesirable from the city's viewpoint if the city already has the problem of trying to find new sources of water. Also, there may be economies of scale in waste treatment so that municipalities can treat wastes at a lower cost than firms can treat them individually, especially for certain types of wastes. It may, therefore, be more efficient socially for the cities to offer municipal treatment of strong wastes at the marginal cost. In addition, it is the definition of NW_i that makes waste dilution an alternative. Another approach would be to have the sewer charges on volume of wastewater and pounds of wastes completely independent, i.e., charge industrial firms for water and have a separate charge for the pounds of BOD (this is equivalent to defining normal wastes as zero ppm of BOD). There would no longer be any advantage to waste dilution. Cities have apparently paid little attention to this particular aspect of the surcharge structure, and one of the conclusions drawn here is that the defined normal waste strength may be an important control variable, especially for some industries.

---

[8] If the firm has to produce a given amount of waste reduction and the alternative of having the city treat part of the wastes as EMT is omitted, then the firm has to increase PR or waste dilution or both. However, if the marginal cost of PR increases as PR increases and the marginal cost of waste dilution decreases as water use increases, which is the case in most cities which meter water, then the firm would use more waste dilution relative to PR.

Another question relates to the waste measurement on which the surcharge is based. The relationships, if any, that exist among different measurements of wastes are not well understood. If, for example, cities were able to achieve the same results with a surcharge on BOD alone as with a surcharge on both BOD and suspended solids, they could avoid the cost associated with sampling and measuring the suspended solids; alternatively some other waste measurement may be more efficient than either BOD or suspended solids.

The estimated response of industries to surcharges has relevance beyond the geographical bounds of cities. Firms not using municipal water and sewer services (firms outside cities) could be expected to respond in essentially the same way as those within the cities if they were faced with the same constraints. This points to the possibility of using charges as a means of controlling waste discharges into public waterways as an addition, or an alternative, to standards and subsidies.

ERRATA

■ In the article "Estimates of the Marginal Rate of Time Preference and Average Risk Aversion of Investors in Electric Utility Shares: 1960–66," by Robert H. Litzenberger and Cherukuri U. Rao, *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), equation (10) on page 273 should read:

$$\rho_i = I + (1 - L_i)S_{\pi_i}. \tag{10}$$

THE BELL JOURNAL OF ECONOMICS AND MANAGEMENT SCIENCE

Editor.
Paul W. MacAvoy, *Massachusetts Institute of Technology*

Associate Editors
Dale W. Jorgenson, *Harvard University*
Harvey J. McMains, *American Telephone and Telegraph Company*
William F. Sharpe, *Stanford University*

Business Manager
Phillip E. Davis, *American Telephone and Telegraph Company*

AUTUMN '72
VOL 3 NO 2.

Table of contents

ARTICLES

✓357 **Capital markets: theory and evidence**
MICHAEL C. JENSEN

399 **Airline regulation and market performance**
THEODORE E. KEELER

✓425 **Quality variation when prices are regulated**
LAWRENCE J. WHITE

437 **Models of the demand for air transportation**
PHILIP K. VERLEGER, JR.

✓458 **Some aspects of the pure theory of corporate finance: bankruptcies and take-overs**
JOSEPH E. STIGLITZ

483 **FCC regulation, monopsony, and network television program costs**
ROBERT W. CRANDALL

✓ 509 **Perfect competition and optimal production decisions under uncertainty**
EUGENE F. FAMA

531 **Nonhomogeneous production functions and applications to telecommunications**
H. D. VINOD

544 **Valuation models in regulation**
BLAINE E. DAVIS AND F. T. SPARROW

568 **An econometric model of the world copper industry**
FRANKLIN M. FISHER, PAUL H. COOTNER AND MARTIN NEIL BAILY

Table of contents (cont'd)

COMMENTS AND REVIEWS

612 **On the use of β in regulatory proceedings**
WILLIAM J. BREEN AND EUGENE M. LERNER

622 **On the use of β in regulatory proceedings: a comment**
STEWART C. MYERS

628 **Albro Martin's Enterprise Denied: a review**
GEORGE W. HILTON

632 **The determination of the allowed rate of return in a formal regulatory hearing**
PAUL L. JOSKOW

645 **Investment behavior in the U.S. electric utility industry, 1949-1968**
U. SANKAR

Capital markets: theory and evidence

Michael C. Jensen

Associate Professor
Graduate School of Management
University of Rochester

This paper is a review of the foundations and current state of mean-variance capital market theory. This work, whose foundations lie in the mean-variance portfolio model of Markowitz, deals with the determination of the prices of capital assets under conditions of uncertainty. The Sharpe-Lintner capital asset pricing model which forms the core of this body of literature is an investigation of the implications of the normative Markowitz model for the equilibrium structure of asset prices. The essential characteristics of these models are reviewed along with the current state of the empirical evidence bearing on them. Many of the recent extensions of the theory are also reviewed and some attempt is made to integrate these extensions with the currently available empirical evidence.

1. Introduction

■ This paper reviews the development of modern capital market theory[1] and the empirical evidence bearing on this theory. There are two main approaches to this problem: the mean-variance models following in the Markowitz tradition, and the state preference models due originally to Arrow and Debreu.[2] Both approaches are generalizations to a world of uncertainty of the work of Irving Fisher on the theory of interest. While the state preference approach is perhaps more general than the mean-variance approach and

---

Michael C. Jensen earned the A.B. degree in economics at Macalester College (1962) and both the M.B A. and the Ph.D. in economics at the University of Chicago, Graduate School of Business (1964 and 1968, respectively). His current research is in the area of finance and economics, with special emphasis on portfolio theory and corporate financial problems

This paper is a revised and extended version of "The Foundations and Current State of Capital Market Theory" [35].

The rate of development of our knowledge of the nature of capital markets at both the theoretical and empirical levels has been extremely rapid in recent times and the author apologizes to the reader for his references to unpublished work and omission of proofs The rapid advances themselves make the task of a survey such as this not only somewhat more difficult but also probably more valuable. The author expresses his appreciation to those authors who have kindly agreed to allow him to reference their results, which in many cases are as yet unpublished, and he apologizes in advance to those authors whose work is not mentioned. This paper is not meant to be an exhaustive survey but only to convey the major thrust of recent work in capital market theory.

The research for this paper was partially supported by the National Science Foundation under grant GS-2964, the University of Rochester Systems Analysis Program under Bureau of Naval Personnel contract number N-00022-70-C-0076, and the Security Trust Company The author wishes to thank Fischer Black, George Benston, David Mayers, Robert Merton, William Meckling, William Sharpe, and Ross Watts, and especially Eugene Fama and John Long for their comments and suggestions.

[1] That is, general equilibrium models of the determination of the prices of capital assets under conditions of uncertainty

[2] See Markowitz [48, 49], Arrow [1], and Debreu [15].

provides an elegant framework for investigating theoretical issues, it is unfortunately difficult to give it empirical content. I restrict attention here to the mean-variance models.[3]

Section 2 provides a brief review of the historical foundations of capital market theory and a statement of the assumptions and implications of the Sharpe-Lintner mean-variance model of asset pricing. A formal derivation of the model and some of its mathematical properties are outlined in the Appendix for the interested reader. Section 3 provides a review of current empirical knowledge about the process determining the structure of asset prices and returns. Section 4 reviews recent theoretical developments in this area and integrates to some extent those developments with our empirical knowledge. Section 5 contains a brief summary and the conclusions.

2. The Markowitz model and the mean-variance asset pricing model

■ The work of Markowitz[4] on portfolio selection resulted in a revolution in the theory of finance and laid the foundation for modern capital market theory. His treatment of investor portfolio selection as a problem of utility maximization under conditions of uncertainty is a pathbreaking contribution. Markowitz deals mainly with the special case in which investor preferences are assumed to be defined over the mean and variance of the probability distribution of single-period portfolio returns, but it is clear that he is aware of the very special nature of these assumptions. In fact, there are very few problems which have received major attention in the literature of this field in the years since the publication of his book that are not at least mentioned there.

Markowitz's treatment of the portfolio problem was almost entirely normative, but other economists began almost immediately to extract positive implications from his approach. The two major directions of these early efforts were (1) Tobin's work[5] utilizing the foundations of portfolio theory to draw implications regarding the demand for cash balances, and (2) the general equilibrium models of asset prices derived by Treynor, Sharpe, Lintner, Mossin, and Fama.[6] Each of these models is an investigation of the implications of the normative Markowitz model for the equilibrium structure of asset prices. They all involve either explicitly or implicitly the following assumptions:

(1) All investors are single-period expected utility of terminal wealth maximizers who choose among alternative portfolios

---

[3] For an excellent exposition of the state-preference approach see Hirshleifer [31, 32, 33]. The "long run growth" model of Markowitz [49], Chap 6, Latane [40], Breiman [10, 11], and Hakansson [29] is another approach to the problem of investment choice under uncertainty See Samuelson [65] for a critique of this approach Since empirical tests of this model are only just beginning to become available and are as yet somewhat incomplete (cf Roll [63]), I shall not consider it here. For criticism of the mean-variance approach see Feldstein [25], Borch [9], the rejoinder by Tobin [71], and Samuelson [66], who provides a justification for the mean-variance approach as an approximation to the exact solution when distributions are "compact "

[4] In [48, 49]

[5] [72]

[6] [74], [67], [42, 43], [56], and [21, 22], respectively

on the basis of mean and variance (or standard deviation) of return.[7]

(2) All investors can borrow or lend an unlimited amount at an exogenously given risk-free rate of interest R_F, and there are no restrictions on short sales of any asset.[8]

(3) All investors have identical subjective estimates of the means, variances, and covariances of return among all assets.

(4) All assets are perfectly divisible and perfectly liquid, i.e., all assets are marketable and there are no transactions costs.

(5) There are no taxes.

(6) All investors are price takers.

(7) The quantities of all assets are given.

Given these assumptions, the previously mentioned authors demonstrate that the equilibrium expected return, $E(\tilde{R}_j)$, on any asset j will be given by

$$E(\tilde{R}_j) = R_F + \lambda \frac{\operatorname{cov}(\tilde{R}_j, \tilde{R}_M)}{\sigma(\tilde{R}_M)}, \tag{1}$$

where R_F is the riskless rate of interest, $\lambda = [E(\tilde{R}_M) - R_F]/\sigma(\tilde{R}_M)$ is the market risk premium per unit of risk, $E(\tilde{R}_M)$ is the expected return on the market portfolio (which consists of an investment in every asset outstanding in proportion to its total value), $\sigma(\tilde{R}_M)$ is the standard deviation of return on the market portfolio, and $\operatorname{cov}(\tilde{R}_j, \tilde{R}_M)$ is the covariance between the return on asset j and the return on the market portfolio. The tildes denote random variables. We now provide a fairly simple derivation of these results and the economic rationale behind them.[9]

☐ **A simple derivation of the model.** Figure 1 gives a geometric presentation of the Markowitz mean-variance model. Letting $\sigma(\tilde{R})$ be the standard deviation of future return, the shaded area in Figure 1 represents all possible combinations of risk and return available from investments in risk-bearing securities. The portfolios lying on the boundary *ABMD* represent the set of mean standard deviation (or mean-variance) efficient portfolios (since they all represent possible investments yielding maximum expected return for given standard deviation and minimum standard deviation for given expected return).

As Tobin has shown,[10] the normality of security returns and the existence of risk aversion on the part of the investor are sufficient to yield a family of positively-sloping convex indifference curves (represented by I_1, I_2, I_3) in the mean standard deviation plane of Figure 1. The shaded area in Figure 1 represents the opportunity set available to the investor in the absence of a riskless asset, and the upper boundary of this set *ABMD* represents the set of efficient

---

[7] Note that representing preferences in terms of mean and standard deviation yields identical results to those obtained with a mean-variance representation.

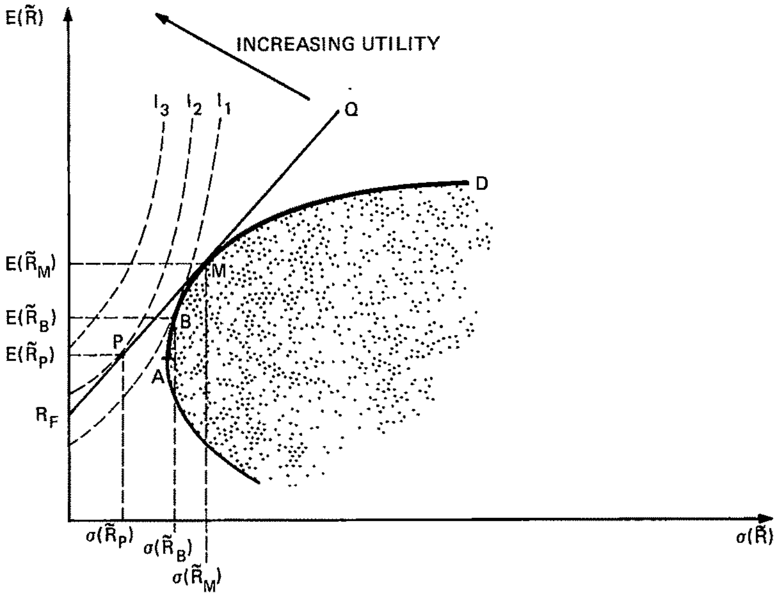
[8] Sharpe [67] is an exception, since he deals with the case where no short sales are allowed.

[9] An alternative and more formal derivation of the results and some additional discussion of the properties of the model are provided in the Appendix.

[10] In [72].

FIGURE 1

THE MAXIMIZATION OF INVESTOR UTILITY GIVEN THE EXISTENCE OF A RISK–FREE ASSET



portfolios.[11] An investor limited only to investments in *risky assets* whose preferences are summarized by the indifference map shown in Figure 1 maximizes his expected utility by investing in portfolio B, which has expected return, $E(\tilde{R}_B)$, standard deviation of return, $\sigma(\tilde{R}_B)$, and yields the investor a level of expected utility denoted by I_1.

Portfolio B portrayed in Figure 1 represents an optimal solution to the portfolio problem only in the case where investment is restricted to risky assets. Let us now assume the existence of a *risk-free asset F*, yielding a certain future return R_F portrayed in Figure 1.[12] An investor faced with the possibility of an investment in such a risk-free asset, as well as in a risky asset (or risky portfolio), can construct a combined portfolio of the two assets which allows him to reach any combination of risk and return lying along a straight line connecting the two assets in the mean standard deviation plane.[13] Clearly, all portfolios lying below point M along $ABMD$ are inefficient, since any point on the line R_FM given by

$$E(\tilde{R}) = R_F + \frac{E(\tilde{R}_M) - R_F}{\sigma(\tilde{R}_M)} \cdot \sigma(\tilde{R}), \qquad \sigma(\tilde{R}) < \sigma(\tilde{R}_M)$$

represents a feasible solution. Thus the investor may distribute his funds between portfolio M and security F such that his combined portfolio, call it P, yields him $E(\tilde{R}_P)$, $\sigma(\tilde{R}_P)$, and maximum utility of $I_2 > I_1$. In addition, if the investor can borrow as well as lend at the riskless rate R_F, the set of feasible portfolios represented by the line R_FM extends beyond point M.

---

[11] See Merton [52] for an analytic discussion of the properties of the efficient set.

[12] Such an instrument might be cash (yielding no positive monetary return), an insured savings account, or a non-coupon-bearing government bond having a maturity date coincident with the investor's horizon date In the latter case, of course, the investor can be assured of realizing the yield to maturity with certainty if he holds the bond to maturity. Since we have assumed the investor will not change his portfolio in the interim period, any intermediate fluctuations in price do not present him with risk. We are ignoring the problems associated with changes in the general price level and shall continue to do so in the remainder of the paper.

[13] Cf. Tobin [72].

Under the assumptions outlined above we can see that *all* investors would hold portfolios lying on the line R_FMQ in Figure 1 and thus would attempt to purchase *only* those assets in portfolio M and the riskless security F.[14] The market for capital assets will therefore be out of equilibrium unless M is the "market portfolio," that is, a portfolio which contains every asset exactly in proportion to that asset's fraction of the total value of all assets. In equilibrium, all investors who select *ex ante* efficient portfolios will have mean standard deviation combinations which lie along the line R_FMQ, their individual location determined by their degree of risk aversion.

Most important, however, is the result that in equilibrium the expected return on any *efficient portfolio* P will be linearly related to the expected return on the market portfolio in the following manner:

$$E(\tilde{R}_P) = R_F + \lambda\sigma(\tilde{R}_P), \tag{1a}$$

where from the geometry of Figure 1 we see that $\lambda = [E(\tilde{R}_M) - R_F]/\sigma(\tilde{R}_M)$. Our problem now is to derive the equation for the equilibrium expected returns on an *individual* asset j.[15] If we let h_j be the fraction of the investor's portfolio invested in the jth asset and $1 - \sum h_j$ the fraction invested in the riskless asset, the return on his portfolio P can be expressed as

$$\tilde{R}_P = \sum h_j\tilde{R}_j + (1 - \sum h_j)R_F.$$

Further, as long as P is an efficient portfolio lying along the line R_FMQ we can express $\tilde{R}_P$ as

$$\tilde{R}_P = \alpha\tilde{R}_M + (1 - \alpha)R_F,$$

where $\tilde{R}_M = \sum X_j\tilde{R}_j$ is the return on the market portfolio, $X_j = h_j/\sum h_i$ is the fraction of the market portfolio invested in asset j, and $\alpha = \sum h_j$. Letting h_j^* represent the optimal investment in the jth asset we can express the expected return and standard deviation of return for any investor's optimal portfolio P (which of course is efficient) as

$$E(\tilde{R}_P) = \sum h_j^* E(\tilde{R}_j) + (1 - \sum h_j^*)R_F = \alpha E(\tilde{R}_M) + (1 - \alpha)R_F$$

$$\sigma(\tilde{R}_P) = [\sum_j \sum_k h_j^* h_k^* \operatorname{cov}(\tilde{R}_j, \tilde{R}_k)]^{1/2} = \alpha\sigma(\tilde{R}_M),$$

where the last equality follows simply from the fact noted above that $\tilde{R}_P$ can be expressed as a linear combination of $\tilde{R}_M$ and R_F.

Now, consider the rate of change of expected return (in equilibrium) in the investor's *optimal* portfolio P as he increases the amount invested in any individual asset h_j by decreasing the amount invested in the riskless asset. To do this we take the derivative of the investor's preference function $U[E(\tilde{R}_P), \sigma(\tilde{R}_P)]$ with respect to

---

[14] This is Tobin's well-known separation theorem. *Ibid.*

[15] I am indebted to John Long for this derivation.

h_j and recognize that at the optimum the rate of change of utility must be zero, so at $h_j = h_j^*$ for all j we have

$$\frac{\partial U}{\partial E(\tilde{R}_P)} \cdot \frac{\partial E(\tilde{R}_P)}{\partial h_j} + \frac{\partial U}{\partial \sigma(\tilde{R}_P)} \cdot \frac{\partial \sigma(\tilde{R}_P)}{\partial h_j} = 0.$$

Rearranging this and recognizing from Figure 1 and the arguments presented above that at the optimum the slope of the investor's indifference curve $-[\partial U/\partial \sigma(\tilde{R}_P)]/[\partial U/\partial E(\tilde{R}_P)]$ must equal λ, the slope of the capital market line, we have

$$\frac{\partial E(\tilde{R}_P)}{\partial h_j} = \lambda \frac{\partial \sigma(\tilde{R}_P)}{\partial h_j}.$$

Substituting for the derivatives obtained from the definitions of $E(\tilde{R}_P)$ and $\sigma(\tilde{R}_P)$ into this expression we obtain

$$E(\tilde{R}_j) - R_F = \lambda \frac{\operatorname{cov}(\tilde{R}_j, \tilde{R}_M)}{\sigma(\tilde{R}_M)}, \tag{1b}$$

where at the optimum,

$$\begin{aligned} \partial \sigma(\tilde{R}_P)/\partial h_j &= \sum_k h_k^* \operatorname{cov}(\tilde{R}_j, \tilde{R}_k)/\sigma(\tilde{R}_P) = \operatorname{cov}(\tilde{R}_j, \sum_k h_k^* \tilde{R}_k)/\sigma(\tilde{R}_P) \\ &= \alpha \operatorname{cov}(\tilde{R}_j, \tilde{R}_M)/\alpha\sigma(\tilde{R}_M), \end{aligned}$$

since $\sum h_k^* \tilde{R}_k = \alpha \tilde{R}_M$. Solving (1b) for $E(\tilde{R}_j)$ then provides the desired result (1).

Thus the equilibrium expected return on any asset[16] is equal to the riskless rate of interest plus a risk premium given by the product of the market risk premium λ and the risk of the jth asset as measured by $\operatorname{cov}(\tilde{R}_j, \tilde{R}_M)/\sigma(\tilde{R}_M)$. Although investors take the variance (or equivalently the standard deviation) of their *portfolio* returns as an appropriate measure of risk, these results imply that the appropriate measure of the risk of any *individual* asset is its covariance with the market portfolio, $\operatorname{cov}(\tilde{R}_j, \tilde{R}_M)$, and not its own variance, $\sigma^2(\tilde{R}_j)$. Diversification can eliminate most of the effects of an asset's own variance on the variance (or standard deviation) of a portfolio[17] but cannot eliminate the effects of an asset's covariances with all other assets in the portfolio. In addition, note

---

[16] Alternatively (as demonstrated in the Appendix) one can express the equilibrium conditions in terms of the price of each asset, V_j, at time 0 as

$$V_j = \left(\frac{1}{1 + R_F}\right)[E(\tilde{D}_j) - \theta \operatorname{cov}(\tilde{D}_j, \tilde{D}_M)], \tag{2}$$

where $E(\tilde{D}_j)$ is the total dollar payoff on asset j at time 1, $\tilde{D}_M = \sum_{i=1}^{N} \tilde{D}_j$ is the total dollar payoff on all assets at time 1, $\operatorname{cov}(\tilde{D}_j, \tilde{D}_M)$ is the covariance between the total dollar returns on asset j and the total dollar returns on all assets, and $\theta = [E(\tilde{D}_M) - (1 + R_F)V_M]/\sigma^2(\tilde{D}_M)$ is the market "price" per unit of risk. Thus the current equilibrium price of any asset is the certainty equivalent of the expected total dollar outcome at time 1 (the term in brackets) discounted back to the present at the riskless rate of interest

[17] Note that the variance $\sigma^2(\tilde{R}_j)$ does affect $\operatorname{cov}(\tilde{R}_j, \tilde{R}_M)$ to some extent since $\operatorname{cov}(\tilde{R}_j, \tilde{R}_M) = \sum X_i \operatorname{cov}(\tilde{R}_j, \tilde{R}_i)$ and one of these N terms is $X_j\sigma^2(\tilde{R}_j)$ Hence the effect is on the order of $1/N$, and given positive dependence among most asset returns the effect of an asset's own variance on its risk will be small relative to the effects of the other $N - 1$ covariances. This is the essence of what are commonly called "portfolio effects."

that cov$(\tilde{R}_j, \tilde{R}_M)$ is proportional to the marginal impact of the jth asset on the standard deviation of the market portfolio, i.e., $\partial\sigma(\tilde{R}_M)/\partial X_j = \text{cov}(\tilde{R}_j, \tilde{R}_M)/\sigma(\tilde{R}_M)$, where X_j is the weight that asset j receives in the market portfolio.[18] Thus rewriting (1) in terms of $\partial\sigma(\tilde{R}_M)/\partial X_j = \text{cov}(\tilde{R}_j, \tilde{R}_M)/\sigma(\tilde{R}_M)$ and defining λ as $[E(\tilde{R}_M) - R_F]/\sigma(\tilde{R}_M)$ we see that in equilibrium the expected return on an asset is linearly related to its marginal contribution to the total risk to be borne by society, $\sigma(\tilde{R}_M)$.

The main result of the original papers in this area is the demonstration that one can derive the individual's demand function for assets, aggregate these demands to obtain the equilibrium prices (or expected returns) of all assets, and then eliminate all the individual utility information to obtain market equilibrium prices (or expected returns) solely as a function of potentially measurable market parameters.[19] Thus the model becomes testable. In addition, the model may be used to address many important practical issues in a number of areas, including valuation theory, the determination of the "cost of capital," corporate investment decisions, governmental cost benefit analysis, and the term structure of interest rates.[20]

3. Empirical tests of the asset pricing model

■ **Evidence from mutual fund returns.** The Sharpe-Lintner asset pricing model has received widespread attention in the literature in the past five years. Most of the early empirical evidence bearing on the model did not represent direct tests, but rather emanated primarily from attempts to use the asset pricing model to derive portfolio performance evaluation models. Treynor, Sharpe, and Jensen all derived such portfolio evaluation models and applied them to the historical evidence on mutual funds.[21] The evidence presented by Sharpe and Jensen[22] indicated that the returns on open-end mutual funds were positively related to the covariance between the fund returns and the returns on a market index used as a proxy for the market portfolio. As such they provided some indications that the model as given by (1) showed potential promise as a description of the process generating the returns on assets (if one was willing to accept the hypothesis that mutual fund managers were unable to systematically select undervalued securities and did not generate excessively large expenses).

☐ **Cross-sectional tests of the model.** Most direct tests of the asset pricing model have been of the cross-sectional form. The average returns on a cross-sectional sample of securities over some time period are regressed against each security's covariance with a market index. Or, more commonly, equation (1) is rewritten as

$$E(\tilde{R}_j) = R_F + [E(\tilde{R}_M) - R_F]\beta_j, \tag{3}$$

---

[18] Cf. Fama [20, 21].

[19] Cf. Appendix.

[20] See the following for applications of the model to these issues. Bailey and Jensen [3], Black and Scholes [7], Fama [20, 21], Hamada [28], Jensen and Long [39], Long [44, 47], Merton [53, 54], Mossin [57], and Roll [64].

[21] See [73], [68], and [36, 37], respectively.

[22] In [68] and [36, 37].

where $\beta_j = \text{cov}(\tilde{R}_j, \tilde{R}_M)/\sigma^2(\tilde{R}_M)$ is what we shall term the *covariance* (or "systematic") risk of the jth asset. Thus the risk of any security is measured relative to the total riskiness of the market portfolio.[23] The procedure is then to estimate the cross-sectional regression

$$\bar{R}_j = \gamma_0 + \gamma_1 \hat{\beta}_j + \tilde{e}_j, \tag{4}$$

where $\hat{\beta}_j$ is obtained from the regression of a time series of individual security returns on an index used as a proxy for the market portfolio. The estimated coefficients $\hat{\gamma}_0$ and $\hat{\gamma}_1$ obtained from the second-stage regression given by (4) are then compared to R_F and $\bar{R}_M - R_F$, respectively, for the time period under consideration. R_F is usually taken to be the yield to maturity of a government bond, with maturity identical to the length of the time period under examination, and $\bar{R}_M$ is the average return on the market index over the period.[24]

The first published direct test of the asset pricing model was that of Douglas[25] who regressed the returns on a large cross-sectional sample of common stocks on their own variance and on their covariance with an index constructed from the sample. For seven separate five-year periods from 1926 to 1960, the average realized return was significantly positively related to the variance of the security's returns over time but not to their covariance with the index of returns.[26] These results appear to be inconsistent with the relation given by (1), since the variance term should have a coefficient of zero.

Douglas also summarizes some unpublished results of Lintner's that also appear to be inconsistent with equation (1). Lintner estimates equation (4) for a cross section of securities over the period 1954–1963. However, he also adds another variable, $\sigma^2(\tilde{u}_j)$ [the variance of the residuals from the time-series regression given by (5)] to the cross-sectional regression. This variance should add nothing to the covariance measure of risk incorporated in β_j and therefore should have a zero coefficient in the regressions. In Lintner's tests it did not. The coefficient on the residual variance was positive and as significant as the β_j term (both having t values greater than six). In addition $\hat{\gamma}_0$ was much greater than R_F and $\hat{\gamma}_1$ much less than $\bar{R}_M - \bar{R}_F$.

---

[23] Note that if the returns on individual securities are linearly related to the market returns, as in

$$\tilde{R}_{jt} = a_j + b_j \tilde{R}_{Mt} + \tilde{u}_{jt}, \tag{5}$$

which is the familiar market (or diagonal model), and if the usual regression assumptions on (5) are satisfied (in particular that $\text{cov}(u_{jt}, u_{it}) = 0, i \neq j$), then $\beta_j = b_j$. Also if (3) is a correct description of the expected returns, then A_j will equal $(1 - \beta_j)R_F$. Cf. Fama [22], Jensen [37], and Black, Jensen and Scholes [8] for discussions of these issues. While the sample estimates of β_j and b_j are always identically equal, this does not imply that the true parameters β_j and b_j are equal They are equal only if the usual regression assumptions on (5) are satisfied, and the equilibrium asset pricing model implies nothing about this issue Thus the "market" model and the asset pricing model, while often used jointly, are different models and the asset pricing model implies nothing about the validity of the market model.

[24] Arithmetic averages of annual, quarterly, or monthly returns have generally been used for both $\bar{R}_j$ and $\bar{R}_M$, but a number of authors have also used average continuously compounded rates as well

[25] [16].

[26] *Ibid*, p. 22

Miller and Scholes[27] review the theory and the Douglas-Lintner evidence and replicate the Lintner results on a different body of data. They confirm his results and then provide a detailed analysis of the possible econometric difficulties involved in estimating these relationships. For example, the biases introduced by various possible misspecifications of the estimating equation include (1) failure to account adequately for the riskless rate of interest, (2) possible nonlinearity in the risk-return relation, and (3) distortions due to heteroscedasticity. However, after explicit examination of these potential biases they conclude that none of them can explain the Douglas-Lintner results, which seem to indicate that an asset's own variance is as important as (or perhaps even more important than) the asset's covariance or portfolio risk in determining its equilibrium price and expected return.

Miller and Scholes also consider the biases which could be introduced by the variables used to approximate returns and risk in the analysis. They examine both theoretically and empirically the possible bias introduced by (1) measurement errors in the risk variable, $\hat{\beta}_j$; (2) the apparent correlation between the residual risk, $\sigma^2(\tilde{u}_j)$, and the covariance risk, $\hat{\beta}_j$; and (3) the utilization of an improper index as a proxy for the returns on the market portfolio. They conclude that measurement errors in $\hat{\beta}_j$ and the observed correlation between $\sigma^2(\tilde{u}_j)$ and $\hat{\beta}_j$ seem to contribute substantially to the Douglas-Lintner results—the former because of the attenuation bias caused by the measurement errors in $\hat{\beta}_j$ and the latter because of the proxying effects of $\sigma^2(\tilde{u}_j)$ for $\hat{\beta}$. However, they conclude that the improper measurement of the market portfolio returns does not seem to be causing substantial problems, and that the other two sources of difficulty are not sufficient to account for all the observed deviations from the models.

Miller and Scholes also investigate the possible difficulties which could be introduced into the empirical analysis by the presence of skewness in the return distributions and the resulting interdependence of sample moments. They show that these skewness effects could cause serious difficulties, and then demonstrate through Monte Carlo techniques that the combined effects of the measurement errors in $\hat{\beta}_j$ and the skewness difficulties could in principle cause the Douglas-Lintner effects, even if the asset pricing model and market model were a completely accurate description of the process generating security returns. While their analysis does not allow us to conclude that the observed empirical results are entirely spurious, they have vividly illustrated many of the econometric difficulties involved in testing the model. As they conclude:[28]

> The check tests we have used are adequate to detect the presence of bias, but too indirect to provide reliable estimates of the "true" relations between risk and return; and certainly not to settle the issue of whether returns really are approximately proportional to nondiversifiable, covariance risk as certain interesting special-case versions of the mean-variance model would suggest. Hopefully, however, more rapid progress in devising test procedures capable of answering these questions can be made, now that attention has been called to some of the hidden minefields along the way.

---

[27] In [55].

[28] *Ibid.*

☐ **Time-series tests of the model.** Black, Jensen, and Scholes (B-J-S)[29] present some additional tests of the Sharpe-Lintner asset pricing model. Most previous studies used cross-sectional tests of a form similar to (4). B-J-S derive a time-series test using procedures similar to those employed by Jensen[30] in the evaluation of mutual fund performance. They argue that if the market model and the asset pricing model are both valid the *ex post* returns on securities will be generated by[31]

$$\tilde{R}_{jt} = R_{Ft}(1 - \beta_j) + \beta_j \tilde{R}_{Mt} + \tilde{e}_{jt}. \qquad (6)$$

If we subtract R_{Ft} from both sides of equation (6) and use primes to denote differences between the return on any asset and the riskless rate we obtain

$$\tilde{R}'_{jt} = \beta_j \tilde{R}'_{Mt} + \tilde{e}_{jt}, \qquad (7)$$

which asserts that the realized risk premium on any asset for any time period is proportional to the realized premium on the market portfolio for that time period plus a random error term with zero expectation. The model can be tested by running the time-series regression given by (7) but allowing a constant term α_j to enter:

$$\tilde{R}'_{jt} = \alpha_j + \beta_j \tilde{R}'_{Mt} + \tilde{e}_{jt}. \qquad (8)$$

If the asset pricing model is valid, the intercept α_j in (8) will be zero. Thus a direct test of the model can be obtained by estimating (9) for a security over some time period and testing to see if α_j is significantly different from zero.[32] While this test is simple, it is inefficient to the extent that it makes use of information on only a single security. In order to solve this problem B-J-S perform their tests on portfolio returns over the period 1931–1965, where the portfolios are constructed so as to maximize the dispersion of their systematic risk.[33] They apply their tests to ten portfolios which contain all securities on the New York Stock Exchange in the

---

[29] In [8]

[30] In [36].

[31] Note that the jump from (3), which is stated in terms of expectations to (6), which is stated in terms of realized observations, is a non-trivial step. The conditions under which this *ex post* formulation of the model are valid are discussed in detail in Fama [22], Jensen [37], and B-J-S [8] As we shall see below some such *ex post* formulation appears to be valid.

[32] Measurement error in the estimate of β_j causes severe difficulties with bias in the cross-sectional tests of the model If $\hat{\beta}_j$ in (4) contains an unbiased measurement error $\tilde{\varepsilon}$, the large sample regression estimate of γ_1 in (4) is

$$\text{plim}\,\hat{\gamma}_1 = \frac{\gamma_1}{1 + \dfrac{\sigma^2(\tilde{\varepsilon})}{S^2(\beta_j)}},$$

where $\sigma^2(\tilde{\varepsilon})$ is the measurement error variance and $S^2(\beta_j)$ is the cross-sectional sample variance of the true risk parameters, β_j Thus even for large samples, as long as $\sigma^2(\tilde{\varepsilon})$ is positive, the estimated coefficient, $\hat{\gamma}_1$, will be biased toward zero and $\hat{\gamma}_0$ will be biased away from its true value, R_F However, this measurement error in β_j introduces no bias whatsoever in the time-series test obtained from (8).

[33] As B-J-S point out, it is important to construct the portfolios in a manner which avoids the introduction of selection bias This can be accomplished by simply avoiding the use of any data for constructing the portfolios which enters into the calculation of the portfolio returns used in the test regressions.

period 1931–1965. The results indicate that the α's are non-zero and are directly related to the risk level, β. Low risk (i.e., low β) securities earn significantly more, on the average, than that predicted by the model ($\alpha > 0$) and high risk securities earn significantly less, on the average, than that predicted by the model ($\alpha < 0$).[34] These results are consistent with those found by Douglas-Lintner and Miller-Scholes regarding the slope and intercept of the cross sectional relationship between average returns and β, but there is little indication that the residual variance $\sigma^2(\tilde{e})$ contributes much to an explanation of the mean portfolio returns.[35] While the evidence indicates that the risk parameters, β_j, are fairly stationary through time there is substantial indication that the intercepts (the α's) are not; a point to which we shall return below. All in all this evidence seems to indicate that the Sharpe-Lintner model *in its most elementary form* does not provide an adequate description of the structure of security returns.

☐ **A two-factor random coefficient model.** B-J-S go on to demonstrate that the process generating the returns on individual securities seems to be well represented by a two-factor model of the form

$$\tilde{R}_{jt} = (1 - \beta_j)\tilde{R}_{Zt} + \beta_j\tilde{R}_{Mt} + \tilde{e}_{jt}, \tag{9}$$

where $\tilde{R}_{Zt}$ represents the return on what they have called the "beta factor," and the other variables are as defined earlier. We consider some alternative (and as yet untested) theoretical justifications for this formulation in Section 4 and concern ourselves here only with the empirical results. Rearranging (9) into the cross-sectional

---

[34] It should be emphasized that these statements apply to the averages and not to the results for every subperiod Indeed, as we shall see, for some subperiods the exact opposite results hold.

[35] The cross-sectional squared correlation between $\bar{R}_j$ and β_j for the period 1931–1965 for the ten portfolios (see Figure 2) was 0.9914. This increased to 0.9916 when the residual variance, $\sigma^2(e_j)$, from (8) was also included in the regression. The estimated regression was

$$\begin{array}{llll} \bar{R}_j = 0.0039 + 0.0102\,\beta_j + 0.9706\,\sigma^2(e_j), & \bar{\beta} = 1.007, s(\beta) = 0.3380 \\ \qquad (0.0004) \quad (0.0004) \qquad (0.3577) & \bar{\sigma}^2(e) = 0.0004, s(\sigma^2(e)) = 0.0004. \end{array}$$

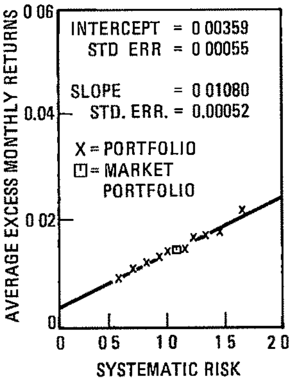
While the coefficient on $\sigma^2(e_j)$ has a t value of 2.71 it explains very little of the mean returns, since the mean of $\sigma^2(\tilde{e}_j)$ is only 0.0004 and its variance is only $(0.0004)^2$. Thus while the variable may be statistically significant, its *economic* significance is virtually nil since $\bar{R}_j$ is on the average equal to 0.0142, or roughly 35 times larger than the average contribution of $\sigma^2(e_j)$ to the portfolio returns. However, from one point of view this may not be the relevant test of the significance of the non-covariance risk in explaining security returns. Unless the market model holds exactly, there is no simple direct relationship between the residual variances of the individual securities and the residual variance of the portfolio in which the securities are contained. Thus one could argue that if there were a significant relationship between the residual variance and return on individual securities this relationship might well become masked by the portfolio grouping procedures and the use of $\sigma^2(e)$ for the portfolio in the cross-sectional regression summarized above. The fact that this might occur only serves to highlight how theoretically implausible are the effects of non-covariance risk since portfolios (in the form of open- and closed-end mutual funds and even diversified companies themselves) are possible investment vehicles, and a system in which these assets are valued differently from single asset claims seems extremely unnatural. We shall see, however, that Fama and MacBeth [23] provide an alternative test of the Douglas-Lintner hypothesis using the individual security residual variances and obtain similar results.

regression form given by (4) we see that the coefficients γ_0 and γ_1 in (4) now become random and are given by $\tilde{\gamma}_{0t} = \tilde{R}_{Zt}$ and $\tilde{\gamma}_{1t} = \tilde{R}_{Mt} - \tilde{R}_{Zt}$, where the $\tilde{R}_{Zt}$ and $\tilde{R}_{Mt}$ are the values of the returns on the beta and the market factors over the holding period of interest.

B-J-S show that if (9) is the process generating asset returns, the cross-sectional tests of the model will provide grossly misleading significance levels of the departures of the data from the model. The reason for this is quite simple. If (9) held where the returns, $\tilde{R}_j$ and $\tilde{R}_M$, are stated in terms of the risk premiums, and if $E(\tilde{R}_Z) = 0$, then (3) or (7) would provide an accurate description of the unconditional expected returns on an asset. Thus a slightly modified version of the Sharpe-Lintner model would be consistent with (9) if $E(\tilde{R}'_Z)$ were zero. However, for a particular cross section $\tilde{R}_{Zt}$ will be a constant. Thus if one runs the cross-sectional regression given by (4) in risk premium form, the intercept γ_0 will be equal to $\tilde{R}'_{Zt}$ (or if the time interval is over more than one period $\bar{R}'_Z$, where $\bar{R}'_Z$ is the mean value of $\tilde{R}'_{Zt}$ over the period). The crucial issue is whether or not the mean, $E(\tilde{R}'_Z)$, is equal to zero, and the regression test of the significance of γ_0 for a single cross section cannot address this question. The usual tests of significance provide a test of whether $\tilde{\gamma}_{0t}$ for a particular cross section is equal to zero and not whether $E(\tilde{\gamma}_0)$ or equivalently $E(\tilde{R}_Z)$ is equal to zero. The usual cross-sectional tests will not incorporate the variability of $\tilde{R}_{Zt}$ through time. The time series tests of B-J-S described above avoid these difficulties and correctly incorporate the variability of the $\tilde{R}_{Zt}$ over time in the significance tests. As we shall see below Fama and MacBeth[36] provide an alternative way of testing the significance of the difference between the average value of γ_{0t} and zero, which is the crucial comparison to be made.

FIGURE 2

AVERAGE EXCESS MONTHLY RETURNS VS SYSTEMATIC RISK FOR TEN PORTFOLIOS AND THE MARKET PORTFOLIO, 1931–1965



B-J-S also show how grouping procedures can eliminate most of the difficulties associated with the biases introduced by measurement errors in the β_j in the cross-sectional tests. Figure 2 summarizes the cross-sectional relationship between the average excess returns, $\bar{R}'_j$, and β's for their ten portfolios (denoted by X's) over the 35-year period 1931–1965. The symbol ◻ denotes the mean excess return, $\bar{R}'_M$, and risk (1.0) for the market portfolio estimated by the average excess returns on all securities on the New York Stock Exchange. If the asset pricing model were valid the relationship should be linear and have an intercept of zero, as equation (7) implies. As can be seen from Figure 2 the relationship is amazingly linear but the intercept is positive and the slope is less than predicted.[37] From the examination of these cross-sectional relationships between risk and return for 17 subperiods of length 24 months and 4 subperiods of length 105 months it appears that on the average the relationship is highly linear, but the intercepts (and slopes) fluctuate randomly from period to period and are often negative. However, as B-J-S argue, these "nonstationarities" are consistent with the return generating process described by the two-factor model[38] of equation (9), which implies that the intercept and slope in the cross-sectional regression given by (4) will be $\bar{R}_Z$ and $[\bar{R}_M - \bar{R}_Z]$, where the bars denote sample means over the time period covered by the cross section. Since $\bar{R}_Z$ will also be a random variable, equation (9) is consistent with the observed empirical results. Because of the existence of sampling error the intercept in the cross-sectional regression given by (4) would not always be identically equal to the riskless rate even if the Sharpe-Lintner model were correct. However, B-J-S argue that the fluctuations in the estimated intercepts $\hat{\gamma}_0$ and slopes $\hat{\gamma}_1$ from cross section to cross section are far greater than that which could be expected from the sampling error associated with the stochastic error term $\tilde{e}_j$ in (4) alone. The standard deviation of this sampling error is far too small to explain the large fluctuations in the coefficients. Thus B-J-S hypothesize that the process generating the data is a two-factor model of the form given by (9).

---

[36] *Ibid.*

[37] It should be equal to $\bar{R}'_M = 0.0142$, but it is 0.0108. The standard errors are given in the upper left-hand corner of Figure 2, but the reader should beware of using them to make significance tests for the reasons outlined above.

B-J-S also provide a tentative procedure for estimating the time series of returns on the beta factor and demonstrate the nonstationarity of its mean $E(\tilde{R}_Z)$ over the period 1931–1965. These findings are consistent with the observed nonstationarity of the intercepts α_j in the time series tests, since for any particular time period α_j in the time series regression, (8) will be equal to $\bar{R}_Z(1 - \beta_j)$.

Jacob[39] in work which is similar in several respects to that of B-J-S also documents the nonstationarity of the risk-return relationship and the inadequacy of the simple form of the Sharpe-Lintner model. Her data include 593 securities listed on the NYSE and covers the period 1946–1965. The results reported by Friend and Blume[40] on the behavior of measures of portfolio performance also appear to be consistent with the B-J-S results and the two-factor model.

□ **A four-factor random coefficient model.** Fama and MacBeth (F-M),[41] in a paper extending the B-J-S work, have found that a four-factor random coefficient model of the form

$$\tilde{R}_{jt} = \tilde{\gamma}_{0t} + \tilde{\gamma}_{1t}\beta_j + \tilde{\gamma}_{2t}\overline{\beta_j^2} + \tilde{\gamma}_{3t}\overline{\sigma_j(u)} + \tilde{\varepsilon}_{jt} \qquad (10)$$

seems to fit the data better than the simple two-factor linear model postulated in the B-J-S study. β_j is as defined before, $\tilde{\gamma}_{0t}$ plays the role of $\tilde{R}_{Zt}$ in the B-J-S notation, $\overline{\beta_j^2}$ is the average of the β^2 for all individual securities in portfolio j and $\overline{\sigma_j(u)}$ is the average of the residual standard deviations from (5) for all securities in portfolio j. The F-M tests were carried out on 20 portfolios constructed from all securities on the NYSE in the period January 1935–June 1968. Like those of B-J-S, the portfolio construction rules used by F-M were designed to minimize the measurement error bias problems in cross-sectional estimates of the coefficients $\tilde{\gamma}_{0t}$, $\tilde{\gamma}_{1t}$, $\tilde{\gamma}_{2t}$, and $\tilde{\gamma}_{3t}$ in (10). The F-M test procedures are based on an examination of the time series of the coefficients in (10), estimated from cross-sectional regressions for each month in the time interval January 1935–June 1968. They focus on three major implications of the equilibrium properties of the Markowitz mean-variance portfolio model (or in F-M's terms the two-parameter portfolio model):

---

[38] As we shall see below, relaxation of certain of the assumptions of the simple model lends some theoretical support to the two-factor model given by equation (9).

[39] [34].

[40] In [27].

[41] [23].

(H1) The risk-return relationship should be linear.
(H2) No measure of risk in addition to β should be systematically related to expected returns.
(H3) The expected return-risk tradeoff should be positive.

(H1), (H2), and (H3) imply respectively that $E(\tilde{\gamma}_{2t}) = 0$, $E(\tilde{\gamma}_{3t}) = 0$, and $E(\tilde{\gamma}_{1t}) > 0$. F-M also test the implications of the traditional Sharpe-Lintner model which says

(H4) $E(\tilde{\gamma}_{0t}) = R_{Ft}$.

In addition they provide tests of the proposition that each of the period-by-period coefficients are equal to zero, $\tilde{\gamma}_{0t} = \tilde{\gamma}_{1t} = \tilde{\gamma}_{2t} = \tilde{\gamma}_{3t} = 0$, and the implication of capital market efficiency that each of the coefficients must behave as a martingale or fair game through time. The latter requirement derives from the fact that, if they did not behave as a fair game, there would exist trading rules based on the past values of the $\tilde{\gamma}$'s which would yield above normal profits—a violation of the efficient markets hypothesis.[42]

From an examination of the average values and t statistics for the estimated coefficients $\hat{\gamma}$ from (10) for the entire period and for various five year subperiods, F-M conclude that the data are consistent with the three major implications of the mean-variance portfolio model (H1)–(H3). The t statistics are defined as

$$t(\bar{\hat{\gamma}}) = \frac{\bar{\hat{\gamma}}}{s(\hat{\gamma})/\sqrt{n}}, \tag{11}$$

where $\bar{\hat{\gamma}}$ and $s(\hat{\gamma})$ are the average and standard deviation of the time series of monthly coefficients and n is the number of months in the period used to calculate $\bar{\hat{\gamma}}$ and $s(\hat{\gamma})$. The average values of $\hat{\gamma}_2$ and $\hat{\gamma}_3$ are generally small and insignificantly different from zero. F-M also confirm the B-J-S results that $E(\tilde{\gamma}_{0t})$ does not equal R_{Ft} (or in the B-J-S terminology $E(\tilde{R}_{Zt}) \neq R_{Ft}$), and therefore (H4), the simple Sharpe-Lintner hypothesis, is inconsistent with the data.

The F-M tests of the period-by-period values of the $\tilde{\gamma}_{jt}$ indicate, however, that while we cannot reject the hypothesis that $E(\tilde{\gamma}_{2t})$ and $E(\tilde{\gamma}_{3t})$ equal zero, we can reject the hypothesis that $\tilde{\gamma}_{2t}$ and $\tilde{\gamma}_{3t}$ equal zero in each month. Therefore while there are no "systematic" nonlinearities in the risk-return relationship and no "systematic" effects of nonportfolio risk on security returns, such effects do materialize in a random fashion from period to period. That is, the data indicate that $\bar{\beta}_j^2$ and $\overline{\sigma_j(u)}$ do help in explaining the period-by-period returns on securities, although their explanatory power, while significantly different from zero, is small in an absolute sense. The average adjusted coefficient of determination in the 402 monthly cross-sectional regressions from 1/35 to 6/68 increases from .29 to .34 with the addition of these last two factors. Furthermore, knowledge of these effects is of no help to the investor since the coefficients themselves behave as a fair game through time. Thus, the investor can apparently do no better than to act as if the two-factor model suggested by B-J-S is valid.

---

[42] Cf [17].

Unfortunately, while we now have a better idea of the nature of the stochastic process generating asset returns, we as yet have no real theoretical understanding of these effects. We shall argue below that there are a number of possible theoretical justifications for these random elements, but as yet few of them are thoroughly tested. As Fama and MacBeth themselves conclude:

> What we have found . . . is that there are variables in addition to [β_j] that systematically affect period-by-period returns. Some of these omitted variables are apparently related to [β_j^2] and [$\sigma_j(u)$]. But the latter are almost surely proxies since there is no economic rationale for their presence in a stochastic risk-return model. . In sum, our results support the important testable implications of the two-parameter portfolio model. We cannot reject the hypothesis that the pricing of common stocks reflects the attempts of risk-averse investors to hold efficient portfolios. On average there seems to be a positive tradeoff between return and risk, with risk measured from the portfolio viewpoint. In addition, although there are "stochastic nonlinearities" from period-to-period, on average their effects are zero and unpredictably different from zero from one period to the next. Thus we cannot reject the hypothesis that in making a portfolio decision, an investor should assume that the relationship between an asset's portfolio risk and its expected return is linear, as implied by the two-parameter model. We also cannot reject the hypothesis of portfolio theory that no measure of risk, in addition to portfolio risk, systematically affects average returns. Finally, the observed "fair game" properties of the coefficients and residuals of our risk-return regressions are consistent with an efficient capital market—that is, a market where prices of securities fully reflect available information.[43]

4. Extensions of the mean-variance asset pricing model

■ The evidence seems to indicate fairly strongly that the simple version of the asset pricing model as described in Section 2 and the Appendix does not provide an adequate description of the process determining common stock returns. The existence of the second, third, and fourth factors documented by B-J-S and F-M indicate that the nature of the process determining security returns through time is quite complex and not at all well understood at the current time. The mere documentation of the existence of a second factor and stochastic nonlinearities in the absence of any real theoretical understanding of the phenomena leaves us in an unsatisfactory state of affairs. In light of this we turn to a closer examination of the theory in search of results which may provide some additional insight into the nature of these phenomena.

Virtually every one of the seven major assumptions (listed in Section 2) upon which the asset pricing model is constructed violates to some degree the conditions observed in the world around us. In attempts to enrich the theory a number of authors have worked on expanding the model by relaxing the seven major assumptions in various ways. The results have indicated that from a theoretical point of view the theory is fairly robust with respect to violations of these assumptions, since many of them are not crucial for the development of the important results of the model. It also appears that some of the extensions of the theory may have in them the seeds of a more complete understanding of the full complexities of the capital markets.

---

[43] [23].

☐ **The single-period utility of terminal wealth assumption.** The assumption that all investors are single-period expected utility of terminal wealth maximizers is clearly erroneous. However, Fama and Long[44] have shown that one can, with little additional difficulty, include the current consumption decisions of investors in the model. That is, the investor's decision is characterized by the simultaneous choice of consumption level and portfolio composition. In addition, Fama[45] provides an analysis of the multi-period consumption-investment problem and a justification for the single-period utility of consumption and terminal wealth model.

Fama argues that the investor's problem is more accurately stated as the maximization of his expected lifetime utility of consumption, $U_{\tau+1}$, i.e.:

$$E[U_{\tau+1}(c_{1-k}, \ldots, c_1, \ldots, c_{\tau+1} | \beta_{\tau+1})],$$

where $C_t = (c_{1-k}, \ldots, c_1, \ldots, c_t)$ is the dollar value of consumption from the beginning of his life, period $1 - k$, through period t. The consumer dies[46] at the beginning of period $\tau + 1$, $c_{\tau+1}$ is his bequest, and $\beta_{\tau+1}$ is the "state of the world" at $\tau + 1$ which signifies the set of all events that constitute history up to and including time $\tau + 1$. The consumer's problem is to make an optimal consumption-investment decision for period 1 taking into account that decisions must also be made at the beginning of each future period (which, of course, will depend on future events). Using the backward optimization of dynamic programming, Fama obtains the recursive relation for all possible states of the world β_t and wealth levels, w_t, as

$$U_t(C_{t-1}, w_t | \beta_t) = \max_{c_t, H} \int_{\beta_{t+1}} U_{t+1}(C_t, w_{t+1} | \beta_{t+1}) dF_{\beta_t}(\beta_{t+1}), \quad (12)$$

subject to the constraints

$$0 \leq c_t \leq w_t$$

$$\sum_i h_i = w_t - c_t$$

$$H \geq 0,$$

where $w_{t+1} = \sum_i h_i \cdot R_i(\beta_{t+1})$ is the investor's wealth at time $t + 1$, $R_i(\beta_{t+1})$ is the one-period per dollar return on asset i in state β_{t+1}, $H = \{h_1, h_2, \ldots, h_n\}$ is the nonnegative vector of dollar amounts invested in each asset at the beginning of period $t + 1$, 0 is the null vector, and $F_{\beta_t}(\beta_{t+1})$ is the distribution function of β_{t+1} given β_t. Fama assumes that consumption goods and portfolio assets are perfectly divisible, there are no transactions costs, and all consumers are price-takers. The function $U_t(C_{t-1}, w_t | \beta_t)$ is the maximum expected utility of lifetime consumption obtainable if the consumer is in state β_t at time t, his wealth is w_t, his past consumption was C_{t-1}, and he makes optimal consumption-investment decisions at time t and in all future periods.

---

[44] In [19,21] and [45], respectively.

[45] In [19]

[46] As Fama demonstrates, the assumption of known time of death is not crucial to the results. In an earlier version Fama [18] demonstrates that the model can be expanded to allow for the existence of random labor income and for the individual's choice between labor and leisure in each period

Fama proves that if the utility of lifetime consumption function $U_{\tau+1}(C_{\tau+1}|\beta_{\tau+1})$ displays risk aversion, i.e., is monotone increasing and strictly concave in $C_{\tau+1}$ for all $\beta_{\tau+1}$, then for all t the derived utility functions given by (12) will have these properties. When the consumer makes a decision at time t, the past consumption pattern, $\hat{C}_{t-1}$, is known. Thus the decision at t can be based upon the function

$$V_{t+1}(c_t, \tilde{w}_{t+1}|\beta_{t+1}) = U_{t+1}(\hat{C}_{t-1}, c_t, w_{t+1}|\beta_{t+1}), \qquad (13)$$

and since V_{t+1} is monotone increasing and strictly concave in (c_t, w_{t+1}) the function $V_{t+1}(c_t, w_{t+1}|\beta_{t+1})$ has the properties of a risk averter's single-period utility of consumption and terminal wealth function for any given state of the world, β_{t+1}.

Note, however, that V_{t+1} given by (13) does not quite provide a complete justification for the simple asset pricing model, since it allows the investor's utility to be state dependent. Fama argues that such state-dependent utilities can arise from three possible situations: (1) tastes for particular bundles of consumption goods can be state dependent; (2) utilities for given dollars of consumption depend on the particular consumption goods available at each point in time and their prices, either or both of which may be state dependent; and (3) the investment opportunities available in any future period may depend on past events and this will induce state-dependent utilities. To justify the simple asset pricing model in the context of the multi-period problem it is thus sufficient to assume that consumers behave as if the future consumption and investment opportunities are given and that tastes are not state dependent. We can then eliminate β_{t+1} from (13) and characterize the investor's decision as the maximization of the expected value of $V_{t+1}(C_t, w_{t+1})$. Thus even though the consumer must solve a multi-period problem to arrive at his optimal current decisions, Fama's results indicate that if he is risk averse, the consumer's observed market behavior will be "indistinguishable from that of a risk-averse expected utility maximizer who has a one-period horizon."[47] Note that V_{t+1} is a complicated function which depends on tastes for future consumption and on the consumption-investment opportunities that will be available in future periods. Hence, while Fama's results do not provide an immediately useful normative rule for making consumption-investment decisions, they do provide a positive justification for the characterization of investor decisions *as if* they were made on the basis of a risk-averse single-period utility of consumption and terminal wealth function.

---

[47] [19], p 164. Some confusion has arisen on occasion over the generality of Fama's results, since he does not impose the separability conditions necessary for decomposition which Nemhauser [58], p. 34ff, requires for utilization of the dynamic programming solution technique. Long [44] derives the general class of utility functions satisfying these conditions, but, as he points out, they are not necessary in Fama's approach. In fact, while decomposition vastly simplifies the calculations necessary to obtain a solution, it is not necessary to insure the optimality of Fama's dynamic programming approach. See Bellman [4], pp. 83–86, where Fama's approach is discussed, and White [76], p. 23ff. Essentially the issue boils down to one involving the definition of dynamic programming and need not concern us here. Also, it has recently come to my attention that Stigum [70] appears to derive results similar to Fama's for corporate decisions under uncertainty.

In addition, Fama demonstrates that if V_{t+1} is concave in c_t and w_{t+1} and the probability distributions are normal, investors can choose among alternative portfolios solely on the basis of mean and variance of returns.[48] He also demonstrates that this result extends to a situation where all probability distributions are symmetric stable with finite mean. In this case, of course, the objects of choice are mean and a dispersion parameter, not mean and variance, since the variance is undefined for all stable distributions except the normal. Fama also shows that one can derive equilibrium conditions equivalent to those of the mean-variance form of the asset pricing model in the context of symmetric stable distributions with finite mean.

Merton[49] derives the equilibrium conditions in a continuous-time version of the model, in which all investors are expected utility of lifetime consumption maximizers and returns are functions of Weiner processes. We return to his results and discuss their implications in somewhat more detail below.

Finally, in a recent paper[50] Long provides what is perhaps one of the most general treatments of the nature of equilibrium in a nonstate preference discrete time multi-period consumption-investment model to date. He assumes risk aversion on the part of consumers and makes no restrictions on the forms of the probability distributions of returns other than that the second moment must be finite. Not surprisingly, it turns out to be impossible to get a directly testable equation like that of (1) without making some additional assumptions. Long does, however, provide a theoretical analysis of the nature of the equilibrium solution under various conditions and examines the comparative statics properties of the model in detail, including its implications for the term structure of interest rates. Of particular interest is his use of the model to demonstrate that risk aversion on the part of consumers does not imply the existence of positive liquidity premiums in the equilibrium term structure. Indeed, he illustrates an economy made up entirely of risk-averse consumers in which the equilibrium liquidity premiums are negative. The "risk" which determines the size and sign of the liquidity premiums is closely related to our usual notion of portfolio risk and in this case is the covariance between the single-period spot rates and the marginal utility of consumption for the corresponding future periods.

☐ **The existence of riskless borrowing and lending opportunities.** Sharpe and Lintner[51] originally assumed that the investor had available unlimited borrowing or lending opportunities at some exogenously fixed interest rate. Mossin[52] assumed the existence of a given quantity of a riskless asset (one whose future value was known with certainty) whose current price was also exogenously given. Long and Fama[53] have shown that the introduction of the consumption decision into the model along with the existence of a fixed quantity of a riskless asset (with short selling possible) allows the riskless interest rate to be determined endogenously. One plus the riskless rate will be equal in equilibrium to the marginal rate of substitution of future for present consumption. Thus the determination of the riskless rate can be brought into the model without changing any of the other results.

---

[48] Tobin [72] provided the first proof of this proposition, but he ignored the consumption decision, c_t, in his analysis.

[49] In [53, 54].

[50] [44].

[51] In [67] and [42, 43], respectively.

[52] In [56].

The model with no riskless borrowing and lending

In a world where there are no contracts denominated in real magnitudes, the presence of uncertainty regarding the general level of prices in the future precludes the existence of a riskless asset. Hence it is important to know to what extent the results of the asset pricing model are dependent on the assumption of the existence of such an asset. Fama[54] demonstrates that given homogenous expectations, no riskless asset, and the ability to sell short all assets, the equilibrium return on any asset will be linearly related to its systematic (or covariance) risk, β. Of course, in the absence of the existence of a riskless rate of interest we can no longer interpret the intercept (or constant) in the equation so easily as in (1) or (3). These results are consistent with the linearity of the empirical relationship between return and β documented by B-J-S, but they do not give us any way to determine the appropriate values of the slope or intercept γ_1 and γ_0 in the cross-sectional relationship given by (4).

Black[55] has demonstrated that one can obtain an equilibrium relationship between risk and return for all assets in a market in which no riskless asset or borrowing or lending opportunities exist (but there are no restrictions on short selling). He proves that in equilibrium the portfolios of all investors consist of a linear combination of two basic portfolios.[56] While this point was recognized earlier,[57] Black lends empirical content to the proposition by demonstrating that the equilibrium conditions imply that one of these two portfolios can be taken to be the market portfolio, M, and the other a portfolio whose returns have zero covariance with the market portfolio.[58] This separation property derives from the fact that given no constraints on short selling, the entire efficient set of portfolios can be generated by a linear combination of these

---

[53] In [45] and [19, 21], respectively

[54] In [21].

[55] In [5].

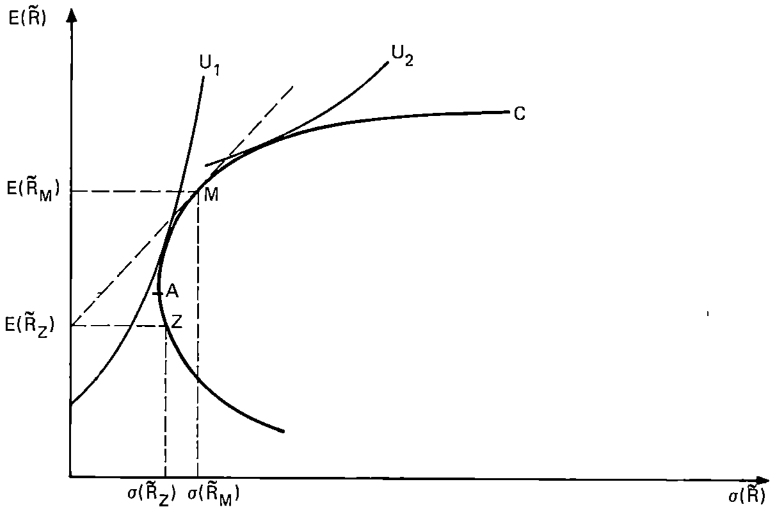
[56] This separation property also holds for the mean-variance portfolio model, as was first pointed out by Tobin [72]. In this case, however, one of the two portfolios was the riskless asset. Cass and Stiglitz [14] proved that the separation property also holds for two risky portfolios for certain classes of utility functions and general return distributions in the absence of a riskless asset. Lintner [41] also investigates the nature of equilibrium in the absence of riskless borrowing or lending for the special case where all investors have preferences which exhibit constant absolute risk aversion and probability distributions are normal.

[57] See Sharpe [69], Chapter 4.

[58] In addition, as Long [46] has pointed out, this portfolio is, of all possible zero covariance portfolios, the one with minimum variance. Black [5] has also shown that the Z portfolio has a covariance with every asset, j, which is proportional to $(1 - \beta_j)$.

FIGURE 3

EQUILIBRIUM PORTFOLIO CHOICE (IN THE ABSENCE OF RISKLESS BORROWING) FOR TWO INVESTORS WITH INDIFFERENCE CURVES U_1 AND U_2



THE MARKET AND MINIMUM VARIANCE ZERO BETA PORTFOLIOS ARE DENOTED BY M AND Z, RESPECTIVELY. THE EFFICIENT SET CONSISTS OF THE LOCUS OF POINTS ON THE SEGMENT AMC

two portfolios,[59] as is demonstrated in Figure 3.[60] The points Z and M represent respectively the standard deviation and expected returns on the zero beta and market portfolios, and since they are uncorrelated, all points on the convex set given by $ZAMC$ can be obtained by various linear combinations of portfolios Z and M. If all funds are invested in portfolio Z, we obtain the risk-return combination given by Z, and as we reduce the fraction invested in Z and invest the remainder in M, we can obtain all points on the curve ZAM. At point M we have nothing invested in Z and hold only M. In order to move beyond M toward C we sell Z short and invest the proceeds in M. Note that all portfolios in the range AMC are efficient in the Markowitz sense. Each investor maximizes his utility by purchasing that combination of Z and M at which his indifference curve between expected return and standard devia-

---

[59] Merton [52] also examines this point In fact the entire frontier of the opportunity set can be generated by linear combinations of any two distinct frontier portfolios.

[60] As pointed out to me by W. F. Sharpe and as shown in Figure 3 the point $E(\tilde{R}_Z)$ must be the intercept of a line tangent to the frontier at point M. To see this, note that the mean and standard deviation, $E(R_P)$ and $\sigma(R_P)$, of a combined investment in Z and M are given by:

$$E(\tilde{R}_P) = \alpha E(\tilde{R}_M) + (1 - \alpha)E(\tilde{R}_Z)$$

$$\sigma(\tilde{R}_P) = [\alpha^2\sigma^2(\tilde{R}_M) + (1 - \alpha)^2\sigma^2(\tilde{R}_Z)]^{1/2},$$

where α is the fraction invested in the market portfolio M. The slope of the tangent line at point M ($\alpha = 1$) is given by

$$\frac{dE(\tilde{R}_P)}{d\sigma(\tilde{R}_P)} = \left.\frac{\partial E(\tilde{R}_P)/\partial\alpha}{\partial\sigma(\tilde{R}_P)/\partial\alpha}\right|_{\alpha=1} = \frac{E(\tilde{R}_M) - E(\tilde{R}_Z)}{\sigma(\tilde{R}_M)},$$

and since the line must pass through the point $[E(\tilde{R}_M), \sigma(\tilde{R}_M)]$ the intercept of the tangent line must be $E(\tilde{R}_Z)$.

tion is just tangent to the efficient set (as is true for indifference curves U_1 and U_2 for hypothetical investors 1 and 2). The equilibrium conditions imply that $E(\tilde{R}_Z)$ must be less than $E(\tilde{R}_M)$[61] and that the holdings of Z must net out to zero for all investors as a whole. Black demonstrates that in equilibrium the expected returns on any asset will be given by

$$E(\tilde{R}_j) = (1 - \beta_j)E(\tilde{R}_Z) + \beta_j E(\tilde{R}_M), \tag{14}$$

where $E(\tilde{R}_Z)$ is the expected return on the zero beta portfolio. Note that the expected returns on all assets are still a linear function of their systematic or covariance risk and that (14) is almost identical to (3) except that $E(\tilde{R}_Z)$ plays the role of R_F, the riskless rate. In addition Mayers[62] demonstrates that if one includes consumption in the model, the expected per dollar returns on the zero beta portfolio will equal a weighted average of the marginal rates of substitution of expected future consumption for present consumption for all individuals; a result which is directly analogous to that obtained for the determination of the riskless rate by Long.[63]

It is also interesting to note that (14) bears a close relationship to the two-factor model suggested by B-J-S. In fact the model given by (9) bears the same relationship to (14) as the market model (5) does to (3), where in (5) A_j is interpreted as equal to $(1 - \beta_j)R_F$. Thus (9) can be interpreted as a two-factor market model, where the expected values of the factors are defined by the equilibrium conditions of the asset pricing model in the context of no riskless borrowing or lending opportunities.

The model with riskless-lending but no borrowing

While the above results are in many ways attractive, one might argue that assumptions have been taken too far. While the nonexistence of perfectly certain real contracts is probably an accurate description of the world, one can argue that the uncertainty introduced by price level changes (at least over relatively short horizons) is infinitesimal. The existence of government bonds which are virtually default-free gives investors the opportunity to lend, if not to borrow, at a rate which is for all practical purposes certain. This suggests investigation of equilibrium in a system with riskless lending opportunities but no riskless borrowing opportunities. Vasicek[64] and Black have demonstrated that the equilibrium risk-return relationships for individual risky securities correspond exactly to that given by equation (14). Their results also imply a piecewise linear relationship between expected return and β for *efficient* portfolios.

Given the existence of riskless lending but no riskless borrowing opportunities, the possible set of standard deviation and expected return combinations available to an investor is portrayed in Figure 4. The efficient set excluding lending opportunities is given by the line segment *ATMC*. The efficient set including optimal utilization

---

[61] A point originally proven by Long [46] and also proven by Vasicek [75].

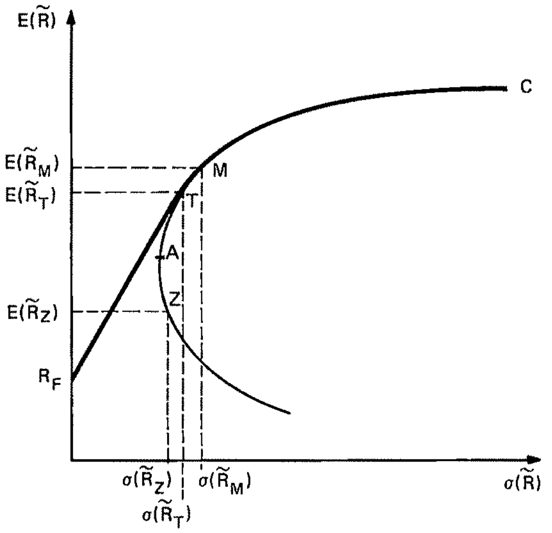
[62] In [51].

[63] In [45].

[64] In [75].

FIGURE 4

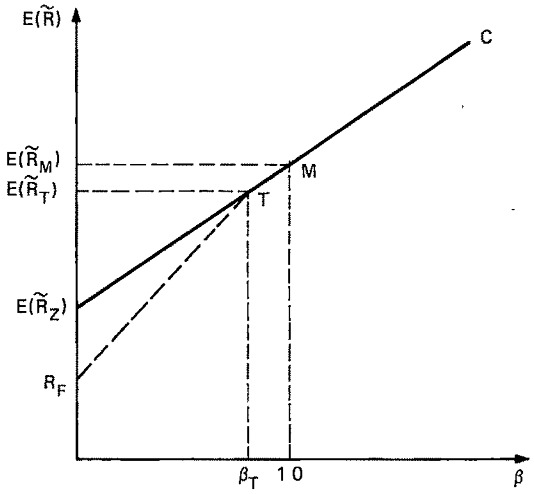
THE LOCUS OF EFFICIENT PORTFOLIOS, $R_F TMC$, WHEN INVESTORS CAN ENGAGE IN RISKLESS LENDING BUT NO RISKLESS BORROWING



THE MARKET PORTFOLIO, ZERO BETA PORTFOLIO, AND TANGENT PORTFOLIO ARE DENOTED BY M, Z, AND T RESPECTIVELY

FIGURE 5

THE RELATIONSHIP BETWEEN EXPECTED RETURN AND SYSTEMATIC RISK, β, WHERE INVESTORS CAN ENGAGE IN RISKLESS LENDING BUT NO RISKLESS BORROWING



ALL EFFICIENT PORTFOLIOS LIE ON THE TWO LINEAR SEGMENTS GIVEN BY $R_F TMC$ AND ALL INDIVIDUAL ASSETS AND INEFFICIENT PORTFOLIOS LIE ON THE STRAIGHT LINE GIVEN BY $E(\tilde{R}_Z)$ TMC

of the lending opportunities is given by $R_F TMC$. Vasicek and Black prove that $R_F \leq E(\tilde{R}_Z) < E(\tilde{R}_M)$ and that in equilibrium all investors will either hold combinations of portfolio T and the riskless asset (if they are on the segment $R_F T$) or combinations of portfolios Z and M (if they are on the segment TMC). Portfolios M and Z are defined exactly as before, i.e., the market portfolio and the zero beta portfolio respectively, and T is a linear combination of these two portfolios. The expected return on every *individual risky asset or portfolio of risky assets* is still given by (14) and lies along the line $E(\tilde{R}_Z)TMC$ in Figure 5. However, unlike the situation for the other models, equation (14) *does not now hold for all efficient portfolios.* In fact the expected return-β relationship for efficient portfolios is now given by the two straight-line segments making up $R_F TMC$ in Figure 5. All efficient portfolios which consist solely of risky assets lie along the portion TMC as in the previous models, *but all efficient portfolios made up of combinations of the riskless asset and portfolio* T now lie along the segment $R_F T$.[65] This result becomes intuitively clear once one notes that if a fraction α is invested in portfolio T and $(1 - \alpha)$ in the riskless asset, the β of the portfolio must equal $\alpha\beta_T$. In addition, the expected returns on the portfolio will equal $\alpha E(\tilde{R}_T) + (1 - \alpha)R_F$, the weighted average of the returns on portfolio T and the riskless asset. Hence for all values of α in the range zero to one the portfolio must lie on

---

[65] These results are entirely consistent with the fact pointed out by Fama [21] that the risk of an asset must be measured in terms of its marginal contribution to the riskiness of some relevant efficient portfolio. If we have two efficient portfolios which are relevant to investor choices, it is not surprising that these results occur. However, the fact that the equilibrium expected returns on all risky assets can still be represented by a simple linear function of β like that given by (14) is interesting.

the line $R_F T$ in Figure 5. However, all individual risky assets or imperfectly diversified portfolios must lie somewhere on the line $E(\tilde{R}_Z)TMC$, and hence those assets and inefficient portfolios with $\beta_j < \beta_T$ will appear to dominate the efficient portfolios lying on the segment $R_F T$ in the mean-beta plane in Figure 5. But of course they do not dominate these efficient portfolios in the mean-standard deviation plane, as illustrated in Figure 4.

Since $E(\tilde{R}_Z)$ must be greater than or equal to R_F, the slope of the line $E(\tilde{R}_Z)TMC$ must be less than or equal to the slope of a hypothetical line drawn from R_F through M, which is the slope given by the simple asset pricing model [cf. equation (3)]. Recall the B-J-S empirical results discussed earlier which were based on an examination of portfolios consisting *entirely of risky assets.* This model predicts that the empirical slope of the cross-sectional return-β relationship found in such a study should be less than that implied by equation (3). If we can assume that for sufficiently long periods of time the sample averages $\bar{R}_M$ and $\bar{R}_Z$ are fairly good proxy measures of the *ex ante* expectations, the B-J-S results are consistent with the predictions of this riskless lending but no riskless borrowing model. Over the period 1931–1965 the empirical slope was less than that predicted by equation (3) [cf. Figure 1], and this also was true for three of the four 105-month subperiods examined.[66]

The model with differential riskless borrowing and lending rates

Brennan[67] derives the market equilibrium conditions assuming that investors can borrow and lend but only at differential rates. He considers two cases: (1) all investors can borrow at a riskless rate r_B and lend at a riskless rate r_L and $r_B > r_L$, and (2) each investor, i, faces different riskless borrowing and lending rates and $r_{Bi} > r_{Li}$. In both cases, Brennan finds that the relationship between an asset's expected return and risk (as measured by β_j) will be linear and identical to (14), the relation obtained by Black for the no riskless borrowing and lending case and by Vasicek for the riskless lending case. He demonstrates that the expected returns on the portfolio Z will be equal to a weighted average of (1) the borrowing rates of borrowers, (2) the lending rates of lenders, and (3) the equivalent riskless rates (marginal rates of substitution) of individuals who neither borrow nor lend. Thus the Black-Vasicek results extend to the general case of differential borrowing and lending rates.

In conclusion, it appears that the relaxation of the assumptions regarding the nature of riskless borrowing and lending opportunities in the original model yields implications which appear to be consistent with some of the observed discrepancies between the empirical results and the simple model documented by B-J-S but do not explain the third and fourth factors involving the nonlinearities and the residual risk components documented by F-M.

---

[66] Interestingly enough, in the last subperiod, April 1957–December 1965, the empirical slope was slightly negative, which is inconsistent with this model. However, this might be explained by the hypothesis that the observed $\bar{R}_M$ and $\bar{R}_Z$ did not adequately represent the *ex ante* expectations, $E(\tilde{R}_M)$ and $E(\tilde{R}_Z)$. Recall also the qualifications noted in Section 2 regarding the interpretation of significance tests in these cross-sectional relationships.

[67] In [12].

☐ **The existence of nonmarketable assets.** The original asset pricing results are based on the assumption that all assets are perfectly liquid. That is, all assets are marketable and there are no transactions costs. Casual observation indicates that this assumption is violated in reality. For example, most investors hold some claims on probability distributions of future income that are nonmarketable; i.e., they cannot sell these claims in current markets. Claims on labor income ("human capital") represent one of the more important of such assets, but there are undoubtedly many more. Nonmarketable claims on the proceeds of assets held in trust and on the proceeds of governmental transfer payment programs such as social security payments are other examples. In addition, there are many physical assets, such as real estate, for which transactions costs are relatively large. Hence it is important to know whether the relaxation of this assumption of perfect liquidity is crucial to any of the results of the model. It is usually quite difficult to analyze such problems in the context of transactions costs (where buying and selling prices differ), because of the discontinuities introduced into the analysis and the dependence of the solution on the initial distribution of resources. However, Mayers[68] demonstrates that for the special case where two kinds of assets exist [perfectly liquid (marketable) and perfectly nonliquid (nonmarketable)] the problem is tractable.

Mayers considers a single-period world in which all investors can have claims on the two types of assets, and derives the individual's demand function for marketable assets (which is a function of the covariance of the returns on the individual's nonmarketable assets with all other assets). He aggregates these demand functions to obtain equilibrium conditions and shows that it is still possible to eliminate information about the individual's utilities and to obtain the equilibrium relationship between the expected return of any asset and its covariance risk in terms of market parameters. His result is

$$E(\tilde{R}_j) = R_F + \lambda \operatorname{cov}(\tilde{R}_j, \tilde{R}_M^* + \tilde{R}_H^*), \tag{15}$$

where

$$\lambda = \frac{E(\tilde{R}_M) - R_F}{\operatorname{cov}(\tilde{R}_M, \tilde{R}_M^* + \tilde{R}_H^*)}$$
$$= \frac{E(\tilde{R}_M) - R_F}{V_M \sigma^2(\tilde{R}_M) + \operatorname{cov}(\tilde{R}_M, \tilde{R}_H^*)}$$

is the "market price" per unit of risk, $\tilde{R}_H^*$ is the *total* dollar return on all nonmarketable assets, V_M is the current total market value of all marketable assets, and all the other variables are as defined earlier except that the asterisks denote total dollar returns. Equation (15) is interesting in several respects, not the least of which is the fact that such a simple equilibrium relationship can be obtained even though investors face widely different efficient sets for their total portfolio opportunities (considering the total probability distribution of returns on marketable and nonmarketable assets). Mayers demonstrates that even with homogeneous expectations, investors will hold widely different portfolios of marketable assets if the nature of the probability distributions on their nonmarketable assets differs. One of the unappealing implications of the simple asset pricing model is that the portfolios of risky assets held by all investors are identical in composition to one another and the market portfolio. This implication, which is clearly inconsistent with the actual portfolio holdings of investors, is not implied by this extended version of the model.

---

[68] In [50]

In addition, note that (15) seems to be a fairly straightforward generalization of (1), where the risk of an individual asset is now measured as the covariance of its per dollar returns with the sum of the returns on the market portfolio and all nonmarketable assets (to allow for the proper weighting of the marketable and nonmarketable assets the returns on each are expressed as total dollar returns instead of per dollar returns). Note that λ can still be interpreted as the market price per unit of risk as measured by the risk premium per unit of risk on the market portfolio. However, the risk of the market portfolio now involves not only its total variance but also its covariance with the returns on nonmarketable assets.

Mayers shows that (15) is equivalent to the equilibrium risk-return relationship that can be derived under the assumption that all assets are marketable when the value and returns of only a subset of these assets (say, those securities listed on the New York Stock Exchange) are measured (or measurable). Under these conditions, equation (15) gives the risk-return relationship for all observed assets in terms of the risk premium on the *observed* market portfolio; i.e., we can interpret M in (15) as the observed market portfolio which is only a subset of the true market portfolio and H as the set of nonobserved, but marketable, assets. Thus the structure of asset returns given the existence of nonmarketable assets is identical to that which would be found if all assets were actually marketable but we could obtain only an incomplete measurement of the assets in the true market portfolio. In addition, Mayers[69] shows that the essential nature of these results remains unchanged if there are no riskless borrowing or lending opportunities, except that $E(\tilde{R}_Z)$ plays the role of R_F in the solutions where $\tilde{R}_Z$ is defined as the return on a portfolio which has zero covariance with the *sum* of the total dollar returns on the market portfolio and the total dollar returns on all nonmarketable assets.

Mayers' results indicate that the basic implications of the theory are not weakened in any major respect by the existence of nonmarketable assets and in fact are strengthened, since the expanded model then does not imply that all investors hold identical proportions of all assets in existence. In addition, his results for the case in which some assets transactions costs are infinite provide some indication that the essential results of the model may hold in the context of finite transactions costs; but of course this conjecture remains to be documented. Mayers' results also imply that the problems introduced by the existence of nonmarketable assets are a generalization of those introduced by the incomplete measurement of $\tilde{R}_M$, and hence his work provides some additional insight into the nature of an ideal index of $\tilde{R}_M$ and the appropriate utilization of imperfect indexes in empirical work. It may be that the effects of nonmarketable and omitted assets will explain the third and fourth factors documented by Fama and MacBeth, but empirical evidence on this issue is at present incomplete.

---

[69] In [51].

☐ **Differential tax rates on dividends and capital gains.** The assumption that there are no taxes is obviously false as a description of reality. Brennan[70] demonstrates under the assumptions of the simple model that the equilibrium price of an asset still can be expressed as a linear function of its systematic risk β_j even when investors face differential tax rates on dividends and capital gains if dividend yields are perfectly certain. He assumes that interest and dividend receipts are taxed at the personal income tax rate and that interest payments are tax deductible. His result is somewhat more complex than equation (3):

$$E(\tilde{R}_j) = T_2 R_F + [E(\tilde{R}_M) - T_1 \delta_M - T_2 R_F]\beta_j + T_1 \delta_j. \tag{16}$$

The major variables are defined exactly as before, and $\delta_j = D_j/V_j$, $\delta_M = D_M/V_M$ are respectively the dividend yields on the jth asset and the market portfolio. The other values are

$$T_1 = \frac{T_d - T_g}{1 - T_g} \tag{17}$$

$$T_2 = \frac{1 - T_d}{1 - T_g} = 1 - T_1, \tag{18}$$

where T_d and T_g are complicated averages of the marginal tax rates[71] on dividends and capital gains respectively for all investors. If taxes on dividends are greater than the taxes on capital gains, T_1 and T_2 will both be positive. Thus the introduction of differential taxes on capital gains and dividends changes the intercept and slope of the equilibrium risk-return relationship and introduces a new variable, the dividend yield, into the determination of expected returns. However, neither the relevant measure of a security's risk nor the linearity of the risk return trade-off (holding dividend yield constant) for individual assets is altered. Note also that the slope implied by (16) is less than that implied by the simple asset pricing model. In addition, the intercept is now $T_2 R_F$, which will be less than R_F if the capital gains tax rate is less than the dividend rate. Furthermore, (16) implies that the higher is an asset's dividend yield, the higher is its equilibrium before tax expected returns. This seems reasonable since, all other things constant, the after tax expected return would otherwise be *lower* the higher the dividend yield.

---

[70] In [13].

[71] The result holds even when the tax rates are progressive.

Brennan examines the relationship between dividend yield and risk (as measured by β) for 100 portfolios constructed from all securities on the NYSE so as to maximize the dispersion of the dividend yields. He finds that yield and risk are strongly negatively correlated (the squared correlation coefficient was 0.59 over the 20-year period 1946–1965). Brennan also estimates the cross-sectional regression

$$\bar{R}_j = \gamma_0 + \gamma_1 \beta_j + \gamma_2 \delta_j + e_j \qquad (19)$$

for the period 1946–1965 using a sample of 11 portfolios constructed from all the securities on the NYSE by procedures also designed to maximize the spread in the δ_j among the portfolios. The coefficients and t values are:

$$\hat{\gamma}_0 = 0.051 \atop (1.89)$$

$$\hat{\gamma}_1 = 0.068 \atop (3.58)$$

$$\hat{\gamma}_2 = 0.634 \atop (3.69),$$

and the squared correlation coefficient is 0.65. Since the returns in (19) are defined as the excess returns above the riskless rate (estimated from 90-day Bankers Acceptances), $\hat{\gamma}_0$ should be zero and $\hat{\gamma}_1$ should be equal to T_1 [cf. equation (16)]. Using this result and the fact that $T_2 = 1 - T_1$ [from (18)] Brennan calculates the theoretical value of $\gamma_1 = [E(\tilde{R}_M) - T_1\delta_M - T_2R_F]$ to be 0.106, which appears to be insignificantly different from $\hat{\gamma}_1$.[72]

Note that $\hat{\gamma}_2$, which is a complex weighted average of the marginal capital gains and dividend tax rates for all investors, falls between zero and one as it should. Brennan also uses instrumental variables procedures in an attempt to eliminate any biases in these results due to specification error and the proxying effects of δ_j for β_j (which are, as indicated earlier, strongly correlated). The results of these tests are virtually the same. Given his estimate of $T_1 = \hat{\gamma}_2 = 0.610$ from the instrumental variables procedures and Bailey's estimate[73] that the average effective tax rate on capital gains is less than 5 percent, Brennan argues that to a rough approximation the results imply that the weighted average dividend tax rate was 63 percent over this period. He notes that this does not differ greatly from the direct estimates obtained by others for various times during his period, which range from 36 percent to 63 percent.[74] On the basis of his regression estimates, Brennan concludes that tax effects on asset yields are important and that his model fits the observed data better than does the simple model.

Black and Scholes[75] also consider the effects of dividend yields on the returns to common stocks. They argue that there is no

---

[72] However, as noted in Section 2, one must be careful about making literal interpretations of the significance levels obtained from such cross-sectional regressions.

[73] In [2].

[74] Cf. [13], p. 175

[75] In [7]

a priori reason to expect the existence of differential taxes on dividends and capital gains to affect the expected returns of any asset. They argue that firms will find it in their best interest to adjust dividend payments so that there is nothing to be gained at the margin from either a reduction or an increase in the total supply of dividends. That is, if some assets sold at either a premium or at a discount (for equivalent risk levels) simply because they had a high dividend yield, other firms would act to eliminate the premium by increasing their dividends or the affected firms themselves could act to eliminate the discount by decreasing or eliminating their dividends. As long as firms are concerned only with the value of their shares and not with the dividend level per se there can be, in equilibrium, no differential values for equivalently risky assets based solely on a differential dividend yield; adjustment on the supply side will eliminate it. Of course, given that dividends are taxed at a higher rate than capital gains, one must assume that some investors prefer dividends to capital gains in order to explain the payment of dividends by any corporation (ignoring constraints imposed by the Internal Revenue Service on the retention of earnings). Brennan does not confront this issue; instead he assumes the supply of dividends is fixed and thus prices need not be equated across differential yield categories by adjustments on the supply side.

Black and Scholes test for the effects of dividends on security prices by using portfolios constructed by ranking securities on dividend yield and then, within a given yield class, ranking all securities on the basis of β_j. They define the yield classes such that 20 percent of all securities on the NYSE in the period 1947–1966 are contained in each class and then divide each of these equivalent yield groups into five portfolios such that the spread on the β's is maximized. When the Brennan cross-sectional regression (19) is run for these 25 portfolios, the coefficients and t values are

$$\hat{\gamma}_0 = \underset{(7.0)}{0.0075}$$

$$\hat{\gamma}_1 = \underset{(4.2)}{0.0028}$$

$$\hat{\gamma}_2 = \underset{(1.3)}{0.0149}$$

and the squared correlation coefficient is 0.46. Thus the intercept and the coefficient on β are "significant" while the coefficient on the dividend yield, δ_j, is much smaller than that obtained by Brennan and "insignificant." To avoid the problems with the interpretation of the significance levels of these cross-sectional tests, Black and Scholes construct a time-series test by combining all the securities into two portfolios by procedures designed to maximize the difference in the dividend yields on the portfolios while constraining the β of each to equal unity. The high yield portfolio has a yield of 6.2 percent and the low yield portfolio, a yield of 4.9 percent. They then run the time-series regression used by B-J-S [equation (8)] to test the asset pricing model. Recall that the intercept α_j in that model should be zero and that B-J-S found that α_j was in fact equal to zero for portfolios with a β of unity but not equal to zero for portfolios with a β different from unity. However, if the dividend yield has an important influence on the total returns of securities, the α on a portfolio which does not have a yield equal to the average value for all securities will not be zero. Thus if dividend yields have an important influence on the returns of assets, as the Brennan model predicts, the intercept α in the time series regression given by (8) should be significantly positive for the high yield portfolio and significantly negative for the low yield portfolio. The coefficients and t values for the period 1947–1966 for the high yield portfolio are:[76]

$$\begin{array}{lll}\tilde{R}'_{jt} = 0.00024 + 1.000\,\tilde{R}'_{Mt}, & & \rho^2 = 0.984 \\ \qquad\;\;(0.72) \qquad (118.3) & & \end{array}$$

and for the low yield portfolio are:

$$\begin{array}{lll}\tilde{R}'_{jt} = -0.00024 + 1.000\,\tilde{R}'_{Mt}, & & \rho^2 = 0.984, \\ \qquad\;\;\;(0.72) \qquad (118.3) & & \end{array}$$

where ρ is the correlation coefficient. While the signs of the intercepts are consistent with the existence of tax effects, each is insignificantly different from zero, and they are insignificantly different from each other (since the standard deviation of the difference between them is equal to 1.44, because the residuals from the two regressions are perfectly negatively correlated by construction).

Black and Scholes also measure the impact of dividend yields on security returns over several subperiods and over the period 1931–1946 with similar results. On the basis of this evidence they state:[77]

> The main conclusion of our analysis is that a dollar of dividends has the same value as a dollar of capital gains in the market There are virtually no differential returns earned by investors who buy high dividend yielding securities or low dividend yielding securities once we control for the crucial risk variable. The demand for dividends is met by corporations who supply dividends and the end result is a market equilibrium in which the dividend factor is insignificant in magnitude

Thus it appears that the Black-Scholes results are inconsistent with the Brennan evidence for reasons that are not readily apparent. Future tests will undoubtedly be required to settle the issue.

☐ **The model in continuous time.** Black and Scholes and Merton,[78] using a continuous time version of the simple form of the asset pricing model, demonstrate that the simple model given by equation (3) holds where the returns are defined over infinitesimally small time intervals. Their results imply that the equilibrium continuously compounded expected rate of return,[79] $E(\tilde{r}_j)$, over any *discrete* interval for any asset j is given by

---

[76] Note that since Black and Scholes used the average excess monthly return on all their securities for the index $\tilde{R}'_M$ the symmetry of the coefficients is a necessary result of the construction of the test and has no economic connotation.

[77] Black and Scholes [7].

[78] In [6] and [53, 54], respectively.

[79] That is, r_j is the natural logarithm of the wealth relative $(P_{jt} + D_{jt})/P_{jt+1}$. Merton [53] also derives a version of the model with no riskless asset in continuous time; his results for instantaneous rates are equivalent to Black's in [5] for discrete time.

$$E(\tilde{r}_j) = \alpha_j + r_F + \beta_j[E(\tilde{r}_M) - r_F], \tag{20}$$

where, except for continuous compounding, all the returns and β_j are defined as before.[80] Note, however, that in addition to the continuous compounding (20) differs from (3) by the addition of the constant α_j, which is given by

$$\alpha_j = [\beta_j \sigma^2(\tilde{r}_M) - \sigma^2(\tilde{r}_j)]/2, \tag{21}$$

where $\sigma^2(\tilde{r}_M)$ and $\sigma^2(\tilde{r}_j)$ are respectively the variances of the continuously compounded returns on the market portfolio and the jth asset over the time interval of observation.[81] Therefore, this model implies that returns should be defined and measured as continuously compounded rates[82] and that the intercept is not necessarily equal to the riskless return. However, given the assumptions of the market model,

$$\sigma^2(\tilde{r}_j) = \beta_j^2 \sigma^2(\tilde{r}_M) + \sigma^2(\tilde{u}_j),$$

we see that

$$\alpha_j = [(\beta_j - \beta_j^2)\sigma^2(\tilde{r}_M) - \sigma^2(\tilde{u}_j)]/2, \tag{22}$$

where u_j can be interpreted as the error term in the continuous time version of the market model given by (5). Substituting this result into (20), we see that the model implies the existence of a much more complicated equilibrium relationship than implied by any of the other models:

$$E(\tilde{r}_j) = r_F + \beta_j[E(\tilde{r}_M) - r_F + (1/2)\sigma^2(\tilde{r}_M)] - \frac{\sigma^2(\tilde{r}_M)}{2}\beta_j^2 - (1/2)\sigma^2(\tilde{u}_j). \tag{23}$$

Moreover, the relation between risk and return is now nonlinear and involves the residual or nonportfolio risk, and is therefore consistent with the third and fourth factors documented by Fama and MacBeth. Note that (23) implies that $\sigma^2(\tilde{u}_j)$ should be negatively related to the level of expected returns; an implication which is inconsistent with the slightly *positive* relationships described earlier[83] and those reported by Lintner and Douglas, Miller and Scholes, and Fama and MacBeth.[84]

To provide a quick and direct test of this model the following cross-sectional regression was run using the data on the ten portfolios utilized in the B-J-S paper discussed in Section 2:

$$\bar{r}_j' = \gamma_0 + \gamma_1 \beta_j + \gamma_2 \beta_j^2 + \gamma_3 \sigma^2(u_j) + e_j, \tag{24}$$

where $\bar{r}_j' = \bar{r}_j - r_F$ is the average continuously compounded excess return on the jth portfolio. For these tests all returns were measured as continuously compounded monthly rates and the β's were calculated using these rates. The results for the 35-year period

---

[80] Cf. Merton [53].

[81] That is, if the returns refer to monthly continuously compounded rates then the $\sigma^2(\tilde{r}_M)$ and $\sigma^2(\tilde{r}_j)$ are variances of the monthly rates

[82] See also Jensen [37].

[83] See note 35 above.

[84] See [16], [55], and [23], respectively.

TABLE 1

REGRESSION TESTS OF THE CONTINUOUS TIME MODEL*

PERIOD	VALUE	γ_0	γ_1	γ_2	γ_3	ρ^2 **	$\bar{r}'_M$	$\sigma^2(\tilde{r}_M)$	$t(\gamma_0)$	$t(\hat{\gamma}_1-\gamma_1)$	$t(\hat{\gamma}_2-\gamma_2)$	$t(\hat{\gamma}_3+1/2)$
1/31–12/65	ESTIMATED	0.00055	0.0161	–0.0066	2.129	0.954	0.0105	0.0070	0 225	0 412	1 107	2.750
	THEORETICAL	0	0.0140	–0.0035	–0.500							
1/31–9/39	ESTIMATED	–0 00884	0.0262	–0 0091	3 483	0.904	0.0108	0 0213	–0.894	0 240	0.143	1.95
	THEORETICAL	0	0 0214	–0 0106	–0 500							
10/39–6/48	ESTIMATED	0 00249	0 0143	–0.0041	0.804	0 899	0.0128	0 0041	0.378	–0 035	0.280	0.713
	THEORETICAL	0	0 0148	–0 0020	–0 500							
7/48–3/57	ESTIMATED	0.00286	0.0149	–0.0055	–3.924	0.734	0.0103	0.0013	0.539	0 339	0.467	0.281
	THEORETICAL	0	0.0110	–0.0006	–0.500							
4/57–12/65	ESTIMATED	0.00092	0 0177	–0.0117	11.198	0.795	0.0082	0.0015	0 106	0 502	1.102	0.938
	THEORETICAL	0	0.0090	–0 0007	–0.500							

*CROSS–SECTIONAL REGRESSION ESTIMATES FOR EQUATION (24) OBTAINED FROM TEN PORTFOLIOS CONTAINING ALL SECURITIES ON THE NYSE IN THE PERIOD 1/31–12/65. ALL RETURNS MEASURED AS CONTINUOUSLY COMPOUNDED MONTHLY RATES SEE EQUATION (23) FOR THE DEFINITION OF THE THEORETICAL VALUES OF THE COEFFICIENTS.

**ρ^2 = SQUARED CORRELATION COEFFICIENT

1931–1965 and for four 105-month subperiods in this interval are presented in Table 1.[85] As can be seen, the coefficients $\hat{\gamma}_1$ for the entire period and the first two subperiods are close to their theoretical values. The values in the last two subperiods are much farther apart. While the t values are also presented in Table 1, their interpretation must be considered in light of the multicollinearity which exists. The variables β_j and β_j^2 are highly correlated (ρ^2 greater than 0.98 in all cases). While this tends to cause the standard errors of estimate of the coefficients to be unreliable, it does not cause bias in the estimates of the coefficients themselves. Also of interest in Table 1 is the fact that $\hat{\gamma}_2$, the coefficient of β_j^2, is about equal to its theoretical value in the first subperiod and roughly one-half its theoretical value for both the whole period and the second subperiod. In the last two subperiods it is considerably different from its theoretical value. Note also that γ_3, the coefficient of the residual variance, $\sigma^2(u_j)$, bears no relationship at all to its hypothesized value. It is positive in 3 out of 4 subperiods but should be equal to -0.5. Multicollinearity problems can be avoided by substituting from (21) into (20), subtracting r_F, and adding $(1/2)\sigma^2(\tilde{r}_j)$ to both sides:

$$E(\tilde{r}'_j) + (1/2)\sigma^2(\tilde{r}_j) = \beta_j[E(\tilde{r}'_M) + (1/2)\sigma^2(\tilde{r}_M)]. \tag{25}$$

Thus if we run the regression

$$\bar{r}_j + (1/2)\sigma^2(\tilde{r}_j) = \gamma_0 + \gamma_1\beta_j + \tilde{e}_j, \tag{26}$$

the coefficients should be given by

$$\gamma_0 = 0 \tag{27a}$$

$$\gamma_1 = [\bar{r}'_M + (1/2)\sigma^2(\tilde{r}_M)]. \tag{27b}$$

---

[85] Some of the numbers in Table 1 are somewhat different from those published in Jensen [38] due to a computer error which caused some of the earlier results to be incorrect.

The regression estimates of (26) are given in Table 2; it is evident that the model does not fit the data. Every coefficient is significantly different from its theoretical value.[86]

While the simple continuous time version of the model is attractive in many ways, these simple tests are not very favorable. However, much more can be done along this line and such work may well prove fruitful.

TABLE 2

CONSTRAINED REGRESSION TEST OF THE CONTINUOUS TIME MODEL*

PERIOD	VALUE	γ_0	γ_1	ρ^2**	$t(\hat{\gamma}_0)$	$t(\hat{\gamma}_1-\gamma_1)$
1/31–12/65	ESTIMATED	0 00299	0 0112	0.975	4.48	–4.68
	THEORETICAL	0	0.0140			
1/31–9/39	ESTIMATED	–0 00888	0.0308	0 950	–3.42	7.82
	THEORETICAL	0	0.0214			
10/39–6/48	ESTIMATED	0.00353	0.0115	0 956	3.86	3.75
	THEORETICAL	0	0.0148			
7/48–3/57	ESTIMATED	0.00743	0.0036	0.617	7.05	7 50
	THEORETICAL	0	0.0111			
4/57–12/65	ESTIMATED	0.01645	–0.0016	0.416	16.46	17.0
	THEORETICAL	0	0 0087			

*REGRESSION ESTIMATES FOR EQUATION (26) OBTAINED FROM TEN PORTFOLIOS CONTAINING ALL SECURITIES ON THE NYSE IN THE PERIOD 1/31–12/65 SEE EQUATION (27) FOR THE DEFINITION OF THE THEORETICAL VALUES OF THE COEFFICIENTS. $\overline{r}'_M$ AND σ^2 $(\tilde{r}_M)$ ARE GIVEN IN TABLE 1.

** ρ^2 = SQUARED CORRELATION COEFFICIENT.

☐ **A three-factor model.** Merton[87] in the context of the continuous time intertemporal asset pricing model derives what we might call a three-factor model. The main point of the Merton paper is to show that any violation of the three conditions used by Fama[88] to justify the single-period utility of terminal wealth model[89] will almost certainly negate the simple form of the model. In particular he concentrates on shifting investment opportunity sets and derives an equilibrium model under the assumption that all such shifts can be characterized by changes in a single variable—the riskless interest rate. Merton demonstrates that under these conditions all investors will hold investments in three portfolios, not two.[90] The three portfolios are: (1) the riskless asset, (2) the market portfolio *M*, and (3) a portfolio (or asset) *N* which is perfectly negatively correlated with changes in the riskless interest rate. Investors will demand shares of portfolio N in order to hedge against the effects of future unforeseen changes in the riskless interest rate. Merton also demonstrates that as long as investor consumption decisions are not independent of the level of r_F, the *instantaneous* expected return on the jth asset, $E(\tilde{r}_j)$, will be given by

---

[86] The 5 percent *t* value for 8 degrees of freedom is 2.31.

[87] [54].

[88] In [19].

[89] That is, consumption preferences, consumption opportunities (for given wealth), and investment opportunities must be independent of the state of the world.

[90] It is also interesting to note that in this model investors will not in general hold mean-variance efficient portfolios. This isn't surprising, since the Fama [19] theorem regarding the reduction of the multi-period model to a single-period mean-variance model discussed above fails here precisely because one of its conditions is violated Allowing the interest rate to change over time causes the investment opportunity set to be state-dependent.

$$E(\tilde{r}_j) = r_F + \lambda_1[E(\tilde{r}_M) - r_F] + \lambda_2[E(\tilde{r}_N) - r_F], \qquad (28)$$

where the "risk" measures are now

$$\lambda_1 = \frac{\beta_{jM} - \beta_{jN}\beta_{NM}}{1 - \rho_{NM}^2},$$

$$\lambda_2 = \frac{\beta_{jN} - \beta_{jM}\beta_{NM}}{1 - \rho_{NM}^2},$$

$$\beta_{jk} = \frac{\operatorname{cov}(\tilde{r}_j, \tilde{r}_k)}{\sigma^2(\tilde{r}_k)},$$

and ρ_{NM} is the correlation between the returns on the market portfolio M and the riskless rate N. Thus even if β_{jM}, the usual definition of systematic risk, were 0, the expected returns on the asset would not be equal to the riskless rate, because the asset returns may be systematically related to changes in the interest rate ($\beta_{jN} > 0$).

There is some evidence that the bond market may not behave in quite the same manner as the equity market and the Merton model may provide some indication of the reasons for this.[91]

Merton argues that in general the sign of λ_2 in (28) will be negative for high beta assets and positive for low beta assets. Furthermore, it can be shown that the sign of $E(\tilde{r}_N) - r_F$ is identical to the sign of $-\partial C/\partial r_F$, where C is the aggregate consumption function. While micro theory tells us nothing about this sign, macro theory usually presumes that aggregate saving is positively related to the interest rate and thus that $E(\tilde{r}_N) - r_F$ should be positive. Merton concludes that this model appears to be consistent with the results found by B-J-S and others that indicate that high beta assets earn less and low beta assets earn more than predicted by the traditional asset pricing model. But until detailed investigation of the empirical implications of the model has been carried out, conclusions on these issues are tentative at best.[92]

☐ **The existence of heterogeneous expectations.** The model as originally derived (and all of the extensions discussed here) are based on the assumption that all investors view the probability distributions on all assets in exactly the same way: an assumption clearly violated in the world. Some have criticized the reality and even the usefulness of the model on these grounds although it is not exceptional in this respect; most of price theory involves essentially the same approach. Indeed, the assumption of perfect certainty, though much stronger than that of homogeneous expectations, has led to many important insights into the nature of markets and resource allocations. The assumption of identical tastes or a "representative consumer" often used in traditional price theory is quite analogous to that of homogeneous expectations, and while such an assumption is clearly erroneous as a description of reality, it has often been found to yield extremely useful insights into complex problems.

---

[91] Roll [62] provides tests of the Sharpe-Lintner model based on taking explicit account of the potential errors involved in using proxy measures of the riskless rate and the market portfolio returns. He applies these tests to a sample of 793 weekly observations on U S Treasury bills beginning in 1949. The results seem to indicate that the asset pricing model does not adequately describe the structure of returns in the Treasury bill market.

[92] Recall that (28) is stated in terms of instantaneous rates, and a more complicated expression applies to discrete time returns as in (20).

A number of authors[93] have examined the relaxation of the assumption of homogeneous expectations and, as one might expect, it in no way plays a critical role regarding the existence of a market equilibrium solution in the two-period model.[94] In addition, Lintner[95] has demonstrated that the relaxation of the assumption eliminates the somewhat unattractive implication that all investors hold identical portfolios of risky assets, but so does the relaxation of some of the other assumptions, as we have seen. Market equilibrium prices can be written in the same form as that of equation (2) but with $E(\tilde{D}_j)$, θ, and $\text{cov}(\tilde{D}_j, \tilde{D}_M)$ interpreted as very complex weighted averages (involving investor utility information) of all investors' expectations.

Unfortunately, these solutions seem to have little empirical content. It appears to be impossible to eliminate the individual utility information from these solutions to obtain equilibrium relationships stated solely in terms of potentially observable market parameters. It is also difficult to obtain closed form expressions for the equilibrium prices in such models, since they require complete knowledge of the marginal rates of substitution of expected return for variance at all points in the mean-variance plane for all investors. By closed form is meant an expression for the equilibrium prices which involves only exogenously given parameters such as probability assessments, quantities of assets, and investor preferences. However, the marginal rates of substitution of expected return for variance depend in general on the equilibrium position of each investor in the mean-variance plane and therefore are themselves functions of the prices of all assets. Since, given heterogeneous expectations, it is difficult if not impossible in most cases to eliminate the utility information from the equilibrium relationships, we cannot obtain prices solely as a function of exogenous parameters.

However, as Lintner has demonstrated,[96] a closed form solution can be obtained in the special case in which all investors' preferences functions, $U^i[E(\tilde{R}), \sigma^2(\tilde{R})]$, can be represented as

$$U^i = E(\tilde{R}) - \alpha_i \sigma^2(\tilde{R}),$$

that is, the case of constant absolute risk aversion in the Pratt-Arrow sense. Since the marginal rate of substitution, $dE(R)/d\sigma^2(R) = -[\partial U/\partial \sigma^2]/[\partial U/\partial E]$, is a constant, α_i, it is obviously independent of market prices. In this case Lintner shows that equation (2) holds and that the market risk premium θ is proportional to the harmonic mean of the α_i values of all investors. In addition, the expected cash flows, $E(\tilde{D}_j)$, and covariances, $\text{cov}(\tilde{D}_j, \tilde{D}_M)$, in (2) can then be interpreted as very complicated averages involving the probability assessments and preferences of all investors for all assets. While such solutions can be obtained, their usefulness is not clear, other than that they provide an illustration of the fact that the structure of equilibrium prices is in many ways similar to that of the simple model.[97]

---

[93] Primarily Lintner [41, 42]. See also Sharpe [69], Appendix D, and Fama [21].

[94] Although as Radner [59, 60] and Long [44] have pointed out, the situation may not be quite this simple in the context of the multi-period problem.

[95] In [41]

[96] *Ibid.*

5. Conclusions

■ The mean-variance asset pricing model has thus far provided a major source of empirically testable propositions regarding the nature of risk and its relation to the equilibrium structure of asset prices. The currently available empirical evidence seems to indicate that the simple version of the asset pricing model[98] does not provide an adequate description of the structure of expected security returns.

The recent evidence presented by Black, Jensen, and Scholes and Fama and MacBeth[99] seems to indicate that the two-factor equation for equilibrium expected returns involving the market factor and the beta factor suggested by B-J-S and later derived theoretically by Black, Vasicek, and Brennan[100] may be an adequate representation of the *unconditional* expected return on assets, even though it is far from a complete specification of the stochastic structure of asset returns. Fama and MacBeth also present evidence which indicates the presence of two additional factors involving β_j^2 and the residual risk $\sigma(u_j)$, both of which have random coefficients whose means are zero. We do not as yet have completely satisfactory explanations for the existence of the beta factor, the stochastic nonlinearities, or the stochastic residual risk effects. However, the evidence seems to be entirely consistent with the three major equilibrium implications of the two parameter portfolio model: (1) The *ex ante* relationship between risk and return is linear, (2) no nonportfolio risk measure is systematically related to *ex ante* expected returns, and (3) market risk premiums are positive. In addition the data also seem to be consistent with market efficiency, since the stochastic coefficients in the monthly cross-sectional regression models of Fama and MacBeth and of Black, Jensen, and Scholes all seem to behave as fair game variables, and therefore do not provide any profitable trading opportunities.

The reality of the model has also been criticized by many. Virtually all of its assumptions have been criticized as inappropriate and suggested as the source of the apparent empirical inadequacies documented earlier by a number of authors.[101] However, as we have seen, most of the assumptions of the model have been shown to be capable of relaxation without destroying the essential nature of the results, and (except for heterogeneous expectations) these models are potentially empirically testable. Thus the mean-variance approach to asset valuation under uncertainty shows promise of a remarkable robustness. Perhaps one of the most promising lines of investigation at this time derives from the two-factor model resulting from the assumption of no riskless borrowing or differential interest rates[102] and the three-factor continuous time model derived by Merton.[103]

---

[97] Lintner [41] also examines the equilibrium solution for the special case of constant absolute risk aversion with heterogeneous expectations where (1) 100 percent margin is required on all short sales, (2) no riskless asset exists, (3) no short sales are allowed, and (4) some investors "have no judgements whatsoever" with respect to some stocks in the market. These solutions are similar in many respects to those outlined above for the simple case of heterogeneous expectations, but they are much more complex.

[98] Cf. Sharpe [67], Lintner [42, 43], Mossin [56], and Fama [22]

[99] In [8] and [23], respectively.

[100] In [5], [75], and [12], respectively

We now have substantial empirical documentation of the process generating the returns on assets and at least several potential theoretical explanations of these empirical results. Although empirical tests of the alternative theoretical explanations are currently in a rather preliminary state, there is little doubt that explicit empirical tests of alternative theories are forthcoming. While we do not yet have a complete solution to the asset pricing problem, in the recent past we have obtained a much better understanding of the nature of the process generating the returns on assets, and it seems likely that the rate of increase in our knowledge in this area will continue at least into the near future.

Appendix

The mean-variance valuation model

■ Consider a world in which all investors (denoted by i) are risk-averse single-period expected utility maximizers whose consumption decisions are made independently of their portfolio decisions. Assume also that all investors' orderings of portfolios can be represented by a utility index of the form $G_i(e_i, v_i)$, where e_i is the expected total cash flow to be generated one period hence by the ith investor's portfolio and v_i is the variance of this cash flow. We also assume that $\partial G_i/\partial e_i > 0$, $\partial G_i/\partial v_i < 0$, all assets are infinitely divisible, and transactions costs and taxes are zero. Define

$$e_i = \sum_j X_{ij}\bar{D}_j - rd_i, \tag{A1}$$

$$v_i = \sum_j \sum_k X_{ij}X_{ik}\sigma_{jk}, \tag{A2}$$

X_{ij} = fraction of the total amount of firm j held by individual i,

$\tilde{D}_j$ = the random total dollar cash flow paid to owners of firm j at time i,

$\bar{D}_j = E(\tilde{D}_j)$,

$$\sigma_{jk} = \begin{cases} \text{var}(\tilde{D}_j), j = k \\ \text{cov}(\tilde{D}_j, \tilde{D}_k), j \neq k \end{cases}$$

$r = (1 + R_F)$, where R_F is the one-period riskless rate of interest at which every consumer can borrow or lend,[104] and

---

[101] Douglas [16], Friend and Blume [27], B-J-S [8], and Miller and Scholes [55].

[102] Cf. Black [5], Vasicek [75], and Brennan [12].

[103] In [53, 54].

d_i = net debt of individual i ($d_i > 0$ implies borrowing, $d_i < 0$ implies lending).

The individual investor's portfolio problem is

$$\max_{X_{ij},\, d_i} G_i(e_i, v_i), \tag{A3}$$

subject to the budget constraint

$$\sum_j X_{ij} V_j - d_i = W_i, \tag{A4}$$

where V_j = total market value of firm j at time 0, and W_i = total wealth of the ith individual at time 0. Under the assumption that all investors can borrow or lend at the rate R_F and that all investors have homogeneous expectations regarding the $\bar{D}_j$ and σ_{jk}, we find the solution to the individual's portfolio problem by forming the Lagrangian

$$L = G_i(e_i, v_i) + \lambda_i\Big[W_i - \sum_j X_{ij} V_j + d_i\Big]. \tag{A5}$$

Differentiating with respect to X_{ij} and d_i and eliminating λ_i provides the optimum conditions for each individual i:

$$\left(\frac{\partial G_i}{\partial e_i}\right)(\bar{D}_j - rV_j) + 2\left(\frac{\partial G_i}{\partial v_i}\right)\sum_k X_{ik}\sigma_{jk} = 0, \quad \text{for all } j \text{ and } i, \tag{A6}$$

or equivalently for all pairs (j, t) we have by rearranging (A6) and dividing the equation for asset j by the equation for asset t:

$$\frac{\bar{D}_j - rV_j}{\bar{D}_t - rV_t} = \frac{\sum_k X_{ik}\sigma_{jk}}{\sum_k X_{ik}\sigma_{tk}}, \qquad \text{for all } i. \tag{A7}$$

Market equilibrium requires that all assets be held, and this provides the condition

$$\sum_i X_{ij} = 1 \qquad \text{for all } j. \tag{A8}$$

Multiplying both sides of (A7) by $\sum_k X_{ik}\sigma_{tk}$, summing over all i, using (A8), and rearranging, we obtain for all pairs (j, t):

$$\frac{\bar{D}_j - rV_j}{\sum_k \sigma_{jk}} = \frac{\bar{D}_t - rV_t}{\sum_k \sigma_{tk}}. \tag{A9}$$

Denoting this common ratio by θ we obtain[105]

$$\theta = \frac{\bar{D}_t - rV_t}{\sum_k \sigma_{tk}}, \qquad \text{for all } t,$$

---

[104] The model as formulated here can also be interpreted as an infinite time horizon model where all probability distributions of single-period cash flows are assumed stationary and r is the single-period riskless rate of interest (assumed identical for all periods).

[105] Since, if $a_1/b_1 = a_2/b_2 = \,..a_n/b_n = Z$, then $\sum_{k=1}^{n} a_i \Big/ \sum_{k=1}^{n} b_i = Z$.

$$= \frac{\sum_t (\bar{D}_t - rV_t)}{\sum_t \sum_k \sigma_{tk}} = \frac{\bar{D}_M - rV_M}{\sigma_M^2}, \tag{A10}$$

where $\bar{D}_M$ = expected total cash flow paid by all firms in the market, V_M = total value of all firms in the market at time 0, and $\sigma_M^2 = \text{var}(\tilde{D}_M)$ = variance of total cash flow paid by all firms.

Substitution from (A10) into (A9) provides an expression for the equilibrium value of the jth firm:

$$V_j = \frac{1}{r}[\bar{D}_j - \theta \operatorname{cov}(\tilde{D}_j, \tilde{D}_M)], \qquad \text{for all } j \tag{A11}$$

where $\operatorname{cov}(\tilde{D}_j, \tilde{D}_M) = \sum_k \sigma_{jk}$. Thus in equilibrium the value of the jth firm is the present value (discounted at the risk-free rate) of the certainty equivalent of the random payment $\tilde{Y}_j$. The certainty equivalent is simply the expected payment $\bar{D}_j$ minus a risk discount given by the product of θ, the price per unit of risk, and the "systematic" risk of the firm given by the covariance of its cash flow with the total cash flows from all other firms (the sum of its own variance and its covariances with all other firms). In addition, defining the expected return on asset j to be $E(\tilde{R}_j) = (\bar{D}_j/V_j) - 1$ and solving (A11) for $E(\tilde{R}_j)$ provides the expression for the equilibrium structure of expected returns on all assets:

$$E(\tilde{R}_j) = R_F + \lambda \operatorname{cov}(\tilde{R}_j, \tilde{R}_M), \tag{A12}$$

where $\lambda = [E(\tilde{R}_M) - R_F]/\sigma^2(\tilde{R}_M)$ is the market risk premium, $E(\tilde{R}_M) = \sum_i z_i E(\tilde{R}_i)$ is the expected return on the market portfolio, $z_i = V_i/V_M$, $\sigma^2(\tilde{R}_M)$ is the variance of return on the market portfolio, and $\operatorname{cov}(\tilde{R}_j, \tilde{R}_M)$ is the covariance between the return on asset j and the return on the market portfolio.

Investor preferences and the market risk premium

☐ An alternative interpretation of θ, the market price per unit of risk can be provided. While equation (A10) defines θ in terms of market parameters this "price" can also be interpreted solely in terms of the weighted average marginal rate of substitution of risk and return for each investor in the economy in equilibrium. To state the value equation in terms of investor preferences, we can use a typical equation from (A6). Rewrite the equation for firm j for the ith investor as:

$$\bar{D}_j - rV_j = 2\left(\frac{\partial e_i}{\partial v_i}\right) \sum_k X_{ik}\sigma_{jk}, \qquad \text{for all } i \text{ and } j, \tag{A13}$$

where $(\partial e_i/\partial v_i) = -(\partial G_i/\partial v_i)/(\partial G_i/\partial e_i) > 0$ is the marginal rate of substitution of expected dollar return for variance for individual i. It can be shown[106] that at equilibrium each investor's portfolio

---

[106] Cf. Mossin [56]. In terms of our analysis this is easily seen by representing the system of equations for the ith investor given by (A6) in matrix notation as

$$\left(\frac{\partial G_i}{\partial e_i}\right)(\bar{D} - rV) + 2\left(\frac{\partial G_i}{\partial v_i}\right)\Sigma X_i = 0, \tag{A6a}$$

where $\bar{D}$ and V are column vectors of expected total cash flows and values respect-contains an identical fraction of the total value of every firm. That is, $X_{ik} = X_{ij} = X_i$ for all k, j, and i. Thus we can rewrite (A13) as

$$\bar{D}_j - rV_j = 2X_i\left(\frac{\partial e_i}{\partial v_i}\right)\sum_k \sigma_{jk}, \qquad \text{for all } i \text{ and } j. \tag{A17}$$

Summing both sides of (A17) over all j, rearranging, and noting that $\sum_j \sum_k \sigma_{jk} = \sigma_M^2$:

$$\frac{\bar{D}_M - rV_M}{\sigma_M^2} = 2X_i\left(\frac{\partial e_i}{\partial v_i}\right), \qquad \text{for all } j. \tag{A18}$$

Summing (A18) over all i and using (A8) and (A10) we see that

$$\theta = \frac{2}{L}\left(\frac{\partial e}{\partial v}\right), \tag{A19}$$

where L is the total number of investors and

$$\left(\frac{\partial e}{\partial v}\right) = \sum_{i=1}^{L} X_i\left(\frac{\partial e_i}{\partial v_i}\right)$$

is the weighted average marginal rate of substitution of return for risk for all investors.[107] (Note that the weight X_i is simply the fractional claim of individual i on the total market cash flow $\tilde{D}_M$.) Thus θ is proportional to the weighted average marginal rate of substitution of expected return for variance for all investors in the economy.

---

ively, $\boldsymbol{\Sigma}$ is the covariance matrix for the random cash flows, and X_i is the column vector of the proportions of each company held by the ith individual Assuming that $\boldsymbol{\Sigma}$ is non-singular and solving for X_i we have

$$X_i = (1/2)\left(\frac{\partial v_i}{\partial e_i}\right)\boldsymbol{\Sigma}^{-1}(\bar{D} - rV). \tag{A14}$$

Since the scalar $(\partial v_i/\partial e_i)$ is the only item on the right-hand side of (A14) which is subscripted by i, the composition of every investor's portfolio is identical up to a scale factor. In addition, each of the elements of the vector X_i is identical; that is, the individual i holds the same fraction of the total value of every firm in existence To see this note that the first conclusion obtained from (A14) indicates

$$X_{ij}/X_{ik} = X_{sj}/X_{sk} \tag{A15}$$

for any two investors i and s and any two securities j and k Multiplying both sides of (A15) by X_{sk}, summing over all s, and using the equilibrium condition (A8) we have

$$\frac{X_{ij}}{X_{ik}} = 1, \tag{A16}$$

and therefore for all individuals i, $X_{ij} = X_{ik} = X_i$ for all assets j and k

[107] Note that $(\overline{\delta e}/\partial v)$ is in equilibrium equal to the harmonic mean of the $(\partial e_i/\partial v_i)$ for all investors. To see this note that

$$2X_i\left(\frac{\partial e_i}{\partial v_i}\right) = \theta \qquad \text{for all } i \tag{A20}$$

Solving this expression for X_i, summing both sides over all i, using (A8), and solving the result for θ we have

$$\theta = \frac{2}{\sum_i \frac{1}{\gamma_i}} = \left(\frac{2}{L}\right)\gamma^*, \tag{A21}$$

where $\gamma_i = (\partial e_i/\partial v_i)$, and $\gamma^* = L/\sum(1/\gamma_i)$ is the harmonic mean of the γ_i's Note also that substituting from (A21) into (A20) for θ and solving for X_i yields $X_i = \gamma^*/\gamma_i L$.

References

1 Arrow, K. J. "Le Role des Valeurs Boursieres pour la Repartition la Meilleur des Risques " *International Colloquium on Econometrics,* 1952, Centre National de la Recherche Scientifique, Paris, 1953, also in English as "The Role of Securities in the Optimal Allocation of Risk-Bearing," *Review of Economic Studies,* Vol. 31, No 2 (April 1964), pp. 91–96

2. Bailey, M. "Capital Gains and Income Taxation," in A. C. Harberger and M. J Bailey, eds., *The Taxation of Income from Capital,* Washington, D. C · The Brookings Institution, 1969.

3. ——— and Jensen, M. C. "Risk and the Discount Rate for Public Investment," in M. C. Jensen, ed., *Studies in the Theory of Capital Markets* [36]

4. Bellman, R *Dynamic Programming.* Princeton: Princeton Univ. Press, 1957.

5 Black, F. "Capital Market Equilibrium with Restricted Borrowing." *Journal of Business,* Vol. 45, No 3 (July 1972), pp. 444–54.

6. ——— and Scholes, M. "Capital Market Equilibrium and the Pricing of Corporate Liabilities." Financial Note No. 16c. Mimeographed. Massachusetts Institute of Technology, September 1970, revised January 1971.

7. ———, "Dividend Yields and Common Stock Returns: A New Methodology." Working Paper No. 488–70, Sloan School of Management, Massachusetts Institute of Technology, September 1970.

8 Black, F., Jensen, M. C., and Scholes, M. "The Capital Asset Pricing Model: Some Empirical Tests," in M. C. Jensen, ed , *Studies in the Theory of Capital Markets* [36]

9. Borch, K. "A Note on Uncertainty and Indifference Curves." *Review of Economic Studies,* Vol. 36, No. 1 (January 1969), pp. 1–4

10 Breiman, L "Investment Policies for Expanding Business Optimal in a Long Run Sense." *Naval Research Logistics Quarterly* (December 1960).

11 ———. "Optimal Gambling Systems for Favorable Games," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley, Calif. · Univ. of California Press, 1961.

12. Brennan, M. J. "Capital Market Equilibrium with Divergent Borrowing and Lending Rates." *Journal of Financial and Quantitative Analysis,* Vol. 6, No 4 (December 1971), pp 1197–1205.

13. ———. "Investor Taxes, Market Equilibrium and Corporate Finance." Unpublished Ph.D. dissertation, Massachusetts Institute Technology, June 1970.

14. Cass, D. and Stiglitz, J E "The Structure of Investor Preference and Asset Returns, and Separability in Portfolio Allocation: A Contribution to the Pure Theory of Mutual Funds." *Journal of Economic Theory,* Vol. 2, No. 2 (June 1970), pp. 122–60.

15. Debreu, G. *Theory of Value.* New York · John Wiley & Sons, Inc., 1959, Chap. 7.

16. Douglas, G. W. "Risk in the Equity Markets: An Empirical Appraisal of Market Efficiency." *Yale Economic Essays,* Vol 9 (Spring 1969), pp 3–45.

17. Fama, E. F "Efficient Capital Markets: A Review of Theory and Empirical Work." *Journal of Finance,* Vol. 25, No. 2 (May 1970), pp. 383–417.

18. ——— "Multiperiod Consumption-Investment Decisions." Report No 6830, Center for Mathematical Studies in Business and Economics, University of Chicago, revised May 1969

19. ——— "Multiperiod Consumption-Investment Decisions." *The American Economic Review,* Vol. 60, No. 1 (March 1970), pp. 163–74.

20. ——— "Perfect Competition and Optimal Production Decisions under Uncertainty." *The Bell Journal of Economics and Management Science,* Vol 3, No. 2 (Autumn 1972), pp. 509–30.

21. ——— "Risk, Return and Equilibrium " *Journal of Political Economy,* Vol. 79, No. 1 (January/February 1971), pp 30–55.

22. ——— "Risk, Return and Equilibrium: Some Clarifying Comments." *Journal of Finance,* Vol. 23, No. 1 (March 1968), pp. 29–40.

23. ——— and MacBeth, J. "Risk, Return and Equilibrium: Empirical Tests." Unpublished manuscript, University of Chicago, July 1972.

24 ——— and Miller, M H. *The Theory of Finance.* New York: Holt, Rinehart & Winston, 1972.

25. Feldstein, M. S "Mean-Variance Analysis in the Theory of Liquidity Preference and Portfolio Selection " *Review of Economic Studies,* Vol. 36, No. 1 (January 1969), pp. 5–12.

26. FISHER, I. *The Theory of Interest.* Rev. ed. New York: Augustus M. Kelley, 1965.
27. FRIEND, I. and BLUME, M. "Measurement of Portfolio Performance under Uncertainty." *The American Economic Review,* Vol 60, No. 4 (September 1970), pp. 561–75.
28. HAMADA, R. S. "Portfolio Analysis, Market Equilibrium and Corporation Finance." *Journal of Finance,* Vol. 24, No. 1 (March 1969), pp. 13–32.
29. HAKANSSON, N. "Capital Growth and the Mean-Variance Approach to Portfolio Selection." *Journal of Financial and Quantitative Analysis,* Vol. 6, No. 5 (January 1971), pp. 517–558.
30. HIRSHLEIFER, J. "Efficient Allocation of Capital in an Uncertain World." *The American Economic Review,* Vol. 54, No. 2 (May 1964), pp. 77–85.
31. ———. "Investment Decisions under Uncertainty: Applications of the State-Preference Approach." *Quarterly Journal of Economics,* Vol. 80, No. 319 (May 1966), pp 252–77
32. ———. "Investment Decision under Uncertainty: Choice-Theoretic Approaches." *Quarterly Journal of Economics,* Vol. 79, No. 317 (November 1965), pp 509–36
33. ———. *Investment, Interest and Capital.* Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1970
34. JACOB, N. "The Measurement of Systematic Risk for Securities and Portfolios: Some Empirical Results." *Journal of Financial and Quantitative Analysis,* Vol. 6, No. 1 (March 1971), pp. 815–34.
35. JENSEN, M C. "The Foundations and Current State of Capital Market Theory," in M. C. Jensen, ed., *Studies in the Theory of Capital Markets* [36].
36. ———. "The Performance of Mutual Funds in the Period 1945–1964." *Journal of Finance,* Vol. 23, No. 2 (May 1968), pp. 389–416.
37. ———. "Risk, the Pricing of Capital Assets, and the Evaluation of Investment Portfolios." *Journal of Business,* Vol. 42, No 2 (April 1969), pp. 167–247.
38. ———, ed. *Studies in the Theory of Capital Markets.* New York. Praeger Publishers, forthcoming.
39. ———, and LONG, J. B., Jr. "Corporate Investment under Uncertainty and Pareto Optimality in the Capital Markets." *The Bell Journal of Economics and Management Science,* Vol. 3, No. 1 (Spring 1972), pp. 151–74.
40. LATANE, H. "Criteria for Choice Among Risky Ventures." *Journal of Political Economy,* Vol. 67, No. 2 (April 1959), pp 144–55.
41. LINTNER, J. "The Aggregation of Investors' Diverse Judgment and Preferences in Purely Competitive Securities Markets." *Journal of Financial and Quantitative Analysis,* Vol. 4, No. 4 (December 1969), pp. 347–400.
42. ———. "Security Prices, Risk, and Maximal Gains from Diversification." *Journal of Finance,* Vol. 20, No. 5 (December 1965), pp. 587–616.
43. ———. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *The Review of Economics and Statistics,* Vol. 47, No. 1 (February 1965), pp. 13–37.
44. LONG, J. B., Jr. "Consumption-Investment Decisions and Equilibrium in the Securities Market," in M. C. Jensen, ed., *Studies in the Theory of Capital Markets* [36].
45. ———. "Corporate Investment under Uncertainty and Pareto Optimality in the Capital Markets." Unpublished manuscript, Graduate School of Management, University of Rochester, August 1970.
46. ———. "Notes on the Black Valuation Model for Risky Securities." Mimeographed. Graduate School of Management, University of Rochester, May 1971.
47. ———. "Wealth, Welfare and the Price of Risk." *Journal of Finance,* Vol. 27, No. 2 (May 1972), pp. 419–33.
48. MARKOWITZ, H M "Portfolio Selection." *Journal of Finance,* Vol. 7, No. 1 (March 1952), pp. 77–91
49. ———. *Portfolio Selection: Efficient Diversification of Investments.* New York: John Wiley & Sons, Inc., 1959
50. MAYERS, D "Non-Marketable Assets and Capital Market Equilibrium under Uncertainty," in M. C. Jensen, ed., *Studies in the Theory of Capital Markets* [36].
51. ———. "Non-Marketable Assets and the Determination of Capital Asset Prices in the Absence of a Riskless Asset." Working Paper No. 7114, Graduate School of Management, University of Rochester, November 1971, forthcoming in the *Journal of Business.*

52. **Merton**, R. C. "An Analytic Derivation of the Efficient Portfolio Frontier." Working Paper No. 493–70, Sloan School of Management, Massachusetts Institute of Technology, October 1970.

53. ——— "A Dynamic General Equilibrium Model of the Asset Market and Its Application to the Pricing of the Capital Structure of the Firm." Working Paper No. 497–70, Sloan School of Management, Massachusetts Institute of Technology, December 1970.

54. ——— "An Intertemporal Capital Asset Pricing Model." Working Paper No. 588–72, Sloan School of Management, Massachusetts Institute of Technology, February 1972, forthcoming in *Econometrica.*

55. **Miller**, M H. and **Scholes**, M. "Rates of Return in Relation to Risk: A Reexamination of Some Recent Findings," in M. C. Jensen, ed., *Studies in the Theory of Capital Markets* [36].

56. **Mossin**, J. "Equilibrium in a Capital Asset Market." *Econometrica,* Vol. 34, No. 4 (October 1966), pp. 768–83.

57. ——— "Security Pricing and Investment Criteria in Competitive Markets." *The American Economic Review,* Vol. 59, No. 5 (December 1969), pp. 749–56

58. **Nemhauser**, G. *Introduction to Dynamic Programming.* New York: John Wiley and Sons, Inc., 1966

59. **Radner**, R. "Existence of Equilibrium Plans, Prices, and Price Expectations in a Sequence of Markets." Technical Report No. 5, Center for Research in Management Science, University of California, Berkeley, June 1970.

60. ——— "Problems in the Theory of Markets under Uncertainty." *The American Economic Review,* Vol. 60, No. 2 (May 1970), pp. 454–60.

61. **Roll**, R. "Assets, Money and Commodity Price Inflation under Uncertainty: Demand Theory." Working Paper No. 48-71-2, Graduate School of Industrial Administration, Carnegie-Mellon University, December 1971.

62 ——— "Bias in Fitting the Sharpe Model to Time Series Data." *Journal of Financial and Quantitative Analysis,* Vol. 4, No. 3 (September 1969), pp. 271–89.

63. ——— "Evidence on the 'Growth Optimum' Model." Working Paper No. 3-71-2, Graduate School of Industrial Administration, Carnegie-Mellon University, July 1971, revised April 1972.

64. ——— "Investment Diversification and Bond Maturity." *Journal of Finance,* Vol. 26, No 1 (March 1971), pp. 51–66.

65. **Samuelson**, P A. "The 'Fallacy' of Maximizing the Geometric Mean in Long Sequences of Investing or Gambling." *Proceedings of the National Academy of Sciences,* Vol. 68 (October 1971).

66. ——— "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means, Variances and Higher Moments." *Review of Economic Studies,* Vol. 38, No 4 (October 1971), pp. 537–42.

67. **Sharpe**, W. F. "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk." *Journal of Finance,* Vol. 19, No 4 (September 1964), pp. 425–42.

68. ——— "Mutual Fund Performance." *Journal of Business,* Vol. 39, No 1, Part 2 (January 1966), pp. 119–38.

69 ——— *Portfolio Theory and Capital Markets.* New York: McGraw-Hill Book Co., 1970.

70. **Stiglum**, B. "Entrepreneurial Choice over Time under Conditions of Uncertainty" *International Economic Review.* Vol. 10, No 3 (October 1969), pp. 426–42.

71. **Tobin**, J "Comment on Borch and Feldstein" *Review of Economic Studies,* Vol. 36, No 1 (January 1969), pp 13–14.

72. ——— "Liquidity Preference as Behavior towards Risk." *Review of Economic Studies,* Vol. 25 (February 1958), pp. 65–85.

73. **Treynor**, J. L. "How to Rate Management of Investment Funds." *Harvard Business Review,* Vol. 43, No. 1 (January/February 1965), pp. 63–75

74. ——— "Toward a Theory of Market Value of Risky Assets" Unpublished manuscript, 1961.

75. **Vasicek**, O. A. "Capital Asset Pricing Model with No Riskless Borrowing." Unpublished manuscript, Wells Fargo Bank, March 1971.

76. **White**, D. J. *Dynamic Programming.* San Francisco: Holden-Day, 1969.

Airline regulation and market performance

Theodore E. Keeler

Assistant Professor of Economics
University of California, Berkeley

In order to determine the effects of domestic airline regulation on the fares and market efficiency of the American air transport industry, it is necessary to know what fares would be if air travel markets were unregulated. To answer this question, a long-run airline cost model is developed and estimated, and it is used to predict hypothetical unregulated (or cost-based) fares for 30 major domestic air travel markets. As a test, the model is used to predict fares on the relatively unregulated California intrastate routes, which it does quite accurately, though the number of such routes observed is small. The results of the study indicate that as of 1968, regulated routes had markups over the estimated unregulated fare ranging from 20 to 95 percent, with a distinct tendency for markups to rise with distance. Crudely updated to 1972, the results indicate current markups of 48 to 84 percent, with less correlation between markup and distance.

1. Introduction

■ For the past 34 years, the Civil Aeronautics Board (CAB) has exerted tight control over interstate airline fares and firm entry. Although the goals and conduct of CAB regulation have been discussed extensively, little has been written on the effects of this regulation on the market performance of the interstate airlines, and there is disagreement among the studies that have considered the question.[1]

Caves, in his classic study of a decade ago, comes to the following conclusion on the performance of the domestic airline industry.

> On the whole, the airlines' record is not bad if compared with unregulated industries of similar seller concentration in the American economy, and it is definitely good by comparison with many consumer goods industries [2]

Theodore E. Keeler received a B.A. degree from Reed College (1967), and the Ph.D. from Massachusetts Institute of Technology (1971).

The author is indebted to M. Adelman, R. Crandall, A. Fisher, D. Gaskins, R. Hall, P. MacAvoy, S. Peck, R. Shiller, and members of the Industrial Organization Workshop at the University of Chicago for many helpful comments and suggestions. Part of the work for this paper was done with the support of a National Science Foundation Graduate Fellowship.

[1] For a complete discussion of the goals of CAB regulation, see Caves [3], pp. 123–70 For a more general and recent discussion of the theory of regulation, see Stigler [25]. A somewhat different view is presented in Posner [22]. The goals of airline regulation will be discussed in this paper in the context of the policy conclusions at the end

[2] Caves [3], p. 428.

Caves bases this result on findings that the airlines have achieved an approximately normal rate of return since the 1930s, have efficient firm size, reasonable selling costs, and reasonable rates of innovation.

If, as Caves asserts, airline performance has been at least equal to that of unregulated oligopolies of similar market structure,[3] it would seem to follow that deregulation would bring about no improvement in market performance (for Caves argues that even in the absence of deregulation, the airline industry would remain a highly concentrated oligopoly). Caves, however, qualifies his results by stating that CAB regulation probably has caused a certain amount of managerial inefficiency, and that deregulation of high-density routes is desirable.[4] Thus his conclusions regarding the effects of CAB regulation on airline market performance are ambiguous.

In a more recent study, Jordan concludes that in the absence of regulation, 1965 interstate trunk airline fares would have been between 32 and 47 percent lower than they were.[5] Jordan reaches these conclusions by comparing fares on the relatively unregulated California intrastate routes (mainly those of Pacific Southwest Airlines (PSA), the most successful California carrier) with fares in the Northwest Corridor, which are regulated by the CAB. Specifically, he estimates hypothetical regulated fares for the California market and then compares them with fares in existence.

It might be highly inaccurate to assume that airlines in the Northwest Corridor face the same cost structure as California airlines. Purvis argues rather persuasively that airport and airways congestion in the Northeast Corridor has raised costs so much that an unregulated, PSA-type airline in 1965 would have charged fares almost as high as Eastern was charging at the time on the New York-Washington route.[6]

From this evidence, then, it would appear that the effects of CAB regulation on airline fares are different for different markets, though past studies shed little light on this. Furthermore, neither Purvis nor Jordan deals with long-haul costs, since no unregulated long-haul carriers exist, and they do little more than speculate as to what fares would be on these routes.

It is the aim of this study to estimate hypothetical unregulated fares on the 30 highest-density trunk airline routes in the United States. To do so accurately requires knowledge of airline cost functions, and of the precise effects of such variables as congestion and trip length on these costs. To estimate such a cost function, it would be preferable to use data for unregulated air carriers. However, for the California intrastate carriers, lack of regulation has also implied lack of data, and even if data were available, the number of observation points would be limited, as would be the range of trip lengths.

The solution to these problems offered here is first to estimate a cost function using interstate airline data, then to adjust it for

---

[3] *Ibid*, p 431.

[4] *Ibid.*, pp 444–49.

[5] See Jordan [10], pp 110–13

[6] See [23].

"efficient" operating practices of the California carriers (such as load factors and seating densities). The cost model is then used to estimate passenger trip costs for each of 30 high-density domestic airline routes, taking account of both stage length and congestion as they affect costs. If they include a normal return on capital, trip costs so estimated should also serve as estimates of competitive, unregulated fares. As a test, the model is used to predict fares on two unregulated California routes. If it succeeds in predicting these fares, then it should also give a good indication as to what unregulated fares would be on other routes.

Before it is possible to estimate a cost model, however, it is necessary to give some consideration to costing methodology and assumptions. Section 2 of this study is devoted to this topic. The costs to be estimated can be divided into direct and indirect expenses; the former are estimated in Section 3 and the latter in Section 4. In Section 5, the relationships estimated are used to predict unregulated fares for the 30 highest-density routes as of 1968. These results are then roughly updated to 1972 to determine the effects of fare and cost changes since 1968. The implications of the results for regulatory policy are considered in the concluding section.

2. Some questions of methodology

■ To find the cost-based fares desired, it is necessary to estimate a long-run airline cost function. Traditionally, such functions have been estimated using cross section data on costs and output for individual firms. The only problem with this approach is that it requires the assumption that all firms in the sample are adjusting factor inputs to minimize costs.[7] In the case of the interstate airlines, with fare regulation and entry restrictions, there is reason to question that assumption.

One method of resolving this problem, developed by this writer for use with the railroad industry and more recently used by Eads, Nerlove, and Raduchel for the local service airlines,[8] entails estimation of a family of short-run cost curves and derivation of a long-run envelope from it. To do this, it is necessary first to determine which factor the firms are most likely not to adjust to minimize costs. If the form of the production function is known, it is then possible to specify and estimate a short-run cost function and to derive the long-run envelope from it.

In the context of the airline industry, Eads, Nerlove, and Raduchel assume that the fixed factor is flight crews (pilots and copilots) and that all other factors are variable. The resulting derived minimum represents what could be achieved if the airlines staffed their flights perfectly.

This procedure is appealing from a theoretical viewpoint, but is it appropriate to estimate the cost functions desired here? What does it mean to minimize costs with respect to pilots' salaries? Very simply, it means that the airline is adjusting its staff so as to make absolutely perfect use of its crews, with not a minute of salaried time wasted. Such perfect staffing is likely to be impossible, if only because of scheduling problems over different routes. Thus to determine the costs for an efficiently staffed airline, it is necessary to look at what the most efficient airlines actually achieve, rather than looking at hypothetical minima which could probably never be realized. Furthermore, it is not clear that airline operations can be realistically aggregated into a Cobb-Douglas production function as Eads, Nerlove, and Raduchel have done. Such aggregation is likely to ignore important details affecting airline costs, such as congestion. These details can be accounted for only with a less aggregative, more engineering-oriented costing approach. The approach taken here will be to break costs down into separate categories to determine output and other variables most likely to affect each category and to estimate the costs using whatever techniques seem most appropriate. Adjustments for potential increases in efficiency (such as higher utilization rates, higher load factors, etc.) will be made separately for each equation.

---

[7] For a complete discussion of the difficulties of statistical costing, see Johnston [9]. See also Meyer and Kraft [16] and Meyer and Straszheim [17], pp. 30–43.

[8] See Keeler [12] and Eads, Nerlove, and Raduchel [6].

Any analysis of long-run costs must include capital rental costs, allowing for both interest and depreciation. For purposes of this analysis, we assume that the productivity of the plant and equipment used is constant over its lifetime. This assumption has been used extensively for calculation of transportation costs, and there is no evidence that it produces grossly inaccurate results.[9] Using this depreciation method yields the following capital rental cost per dollar of gross investment:

$$\rho = \frac{r}{1 - e^{-rL}}$$

where ρ is the rental rate, r is the interest rate, and L is the life of the capital good.

To determine the appropriate interest rate for airline cost calculations, in the absence of any special risk considerations, it would not be inappropriate to use the opportunity cost of capital in the rest of the corporate sector: if, over the long term, the airlines failed to achieve this rate of return, they would probably be driven from business. Caves concludes that for the 30 years up to 1960, the airlines earned almost exactly the opportunity cost of capital in the entire corporate sector, about 7.5 percent.[10] Updating this research through 1966, the present writer found almost identical results: The average rate of return for the corporate sector as a whole for 1960–1966 was 11.7 percent before corporate taxes and 7.8 percent after taxes.[11] For the airlines, the average return over the same period was 7.5 percent after taxes.[12] Since the airlines have managed, over the years, to get adequate capital for expansion with this rate of return (if anything, they have managed to get all too much capital recently), this return is clearly adequate. The higher rates of return advocated by the airlines and the CAB would clearly not appear to be justified. Given the corporate income tax structure and the current amount of leverage in the airline and entire corporate sectors, a 12-percent capital cost should allow for a 7.5-percent return after taxes, plus payment of the corporate income tax.

---

[9] For a previous application of this assumption to the transportation industries, see Peck and Meyer [21].

[10] See [3], p. 394

[11] See Keeler [13], p. 20

[12] *Ibid.*, p. 22.

Because airlines own very little in the way of buildings and longlasting facilities, the lifetime chosen for airline capital goods must consider mainly aircraft and ground equipment. Peck and Meyer suggest that a depreciation period of 6 years would be most appropriate for airline equipment;[13] but much jet equipment built in the early 1960s is still in use and serving efficiently, so 6 years would seem too short. Most airlines today use lives of 9 to 15 years for their equipment.[14] If there is a central tendency among airline depreciation policies, it would be about 12 years. That is the figure assumed here.

3. Estimation of direct airline costs

■ Direct airline costs include expenses for crew salaries, fuel, aircraft maintenance, and aircraft capital investment. The most important determinant of direct airline operating costs is block-hours of operation for each plane type. Block-hours measure the total amount of time elapsed between the time the plane taxies away from one gate and the time the engines are shut off at the destination gate.

Since a certain amount of time is necessary for each takeoff and landing, there is an expense connected with each flight hop that is independent of length. To estimate trip costs, this fixed amount of time is added to travel time (based on distance and the cruising speed of the plane involved) to get total block time. This time is multiplied by the cost per block-hour for the appropriate aircraft. The standard procedure, then, is to graph the estimated average cost (per seat-mile or per plane-mile) against stage length. This approach (or one very similar to it) is used for most studies of direct airline operating costs.[15] Its main weakness is that it ignores the fact that flight time depends on congestion as well as on distance; if congestion substantially increases flight time on some routes, this procedure will introduce a strong downward bias in cost estimates.

These facts suggest the following methodology for estimation of direct operating costs of a reasonably efficient airline. First, obtain estimates of normal or typical operating costs per block-hour for the most efficient available aircraft; these estimates should include a normal airline rate of return on capital. Next, apply the seating densities and load factors that California intrastate carriers (especially Pacific Southwest Airlines (PSA)) have been able to achieve (yielding a cost per passenger-hour). Block times for each route depend somewhat on local conditions and should therefore be based on actual travel times for each route. Hence, feasible block times are derived from schedules for 30 city pairs (with allowance for late arrivals).[16] Combining estimated costs per passenger-hour with trip times yields an estimate of passenger trip costs for each of the 30 routes.

---

[13] See [21].

[14] This range was found by looking at various airline reports in *Moody's Transportation Manual* [18].

[15] This is what the CAB does in costing service on local service routes. See *Local Service Air Carriers' Unit Costs* [30], March 1970 edition, pp. 6–15, for example. This is also very similar to the approach of Caves [3], pp. 65–73, and of Straszheim [26], pp. 85–90.

[16] This approach is basically the same as that used by Purvis [23], although Purvis does not allow for the possibility of late arrivals

In estimating direct airline costs, it is important to disaggregate by aircraft type, since costs differ substantially from plane to plane. Three of the most efficient types available are considered here. They are all second-generation "stretched" aircraft introduced in the late 1960s (there are not yet enough reliable data on the wide-bodied aircraft introduced in the early 1970s for accurate cost estimation). Aircraft types considered here include the Douglas DC8-61, designed for long hauls; the Boeing 727-200, designed for short and medium hauls; and the Douglas DC9-30, designed for short hauls.

Past studies agree that the most important determinant of direct operating costs is block-hours of operation for each aircraft type. When statistical estimation of direct costs was attempted, block-hours was generally the only significant variable. Therefore, we assume operating cost per block-hour to be constant for each aircraft type. Table 1 shows cash cost estimates, in 1968 prices, for the three aircraft types under consideration. For all accounts except maintenance, these costs represent the average cost per block-hour for all domestic trunk operations. Maintenance costs deserve special attention. All three aircraft types were practically new in 1968, and for airlines with deferred maintenance policies (which includes most of them), use of raw maintenance cost data will probably give downward-biased estimates of hourly costs over the full life of the plane. Fortunately, there is a simple solution to this problem. Each of the three stretched aircraft are practically identical to their unstretched parents except for fuselage length.[17] Furthermore, the older aircraft were reasonably well broken in by

TABLE 1

DIRECT AIRCRAFT OPERATING COSTS[a]

EXPENSE CLASS	COST PER BLOCK–HOUR		
	AIRCRAFT TYPE		
	DC8–61	727–200	DC9–30
CASH EXPENSES			
FLYING OPERATIONS	$393	$294	$220
MAINTENANCE	186	167	129
TOTAL CASH EXPENSES	579	461	349
TOTAL CAPITAL EXPENSES[b]	343	308	191
TOTAL DIRECT OPERATING COSTS	922	769	540

[a]COSTS PER AIRCRAFT BLOCK–HOUR ARE IN 1968 PRICES.

[b]FROM TABLE 2

ALL FIGURES ARE AVERAGES FOR THE DOMESTIC TRUNK CARRIERS, AND MAINTENANCE COSTS ARE FOR THE UNSTRETCHED VERSIONS OF THE PLANES, (SEE SECTION 3)

SOURCE CAB, AIRCRAFT OPERATING COST AND PERFORMANCE REPORT, [28].

---

[17] For the Boeing 727–200, for example, 80 percent of all parts, and all engine parts, were interchangeable with the shorter 727–100. See Woolsey [34].

the late 1960s. Hence, the maintenance cost figures used are for the earlier, unstretched parents of the aircraft listed.

As a check on this procedure, consider the 727-200 maintenance costs for Continental Airlines, which has a progressive maintenance policy wherein overhauls are done before they are absolutely necessary, thus making more efficient use of equipment and maintenance crews.[18] For Continental, these costs were $151.09 per block-hour in 1968 and $163.77 in 1969, compared with the figure of $167.28 assumed here (presumably there would be some increase in these costs over time even with progressive maintenance, but evidently not much).

Direct capital costs are computed in Table 2. To derive these costs for each plane type, it is necessary to know the utilization rate as well as interest costs. We use the average utilization rate for each plane type in 1968. One might expect this to vary roughly with stage length, since aircraft on long hauls can be used more of the time. The estimated operating costs will implicitly account for this, however, since each aircraft is assigned to a route for cost calculations on the basis of appropriate stage length.

TABLE 2

DIRECT AIRCRAFT CAPITAL COSTS

	AIRCRAFT TYPE		
	DC8–61	727–200	DC9–30
PURCHASE PRICE	$9,000,000	$6,700,000	$4,000,000
ANNUAL COST AT 12% (12–YEAR LIFE IMPLIES CAPITAL RECOVERY COST=0 1576	1,418,000	1,056,000	630,000
ANNUAL UTILIZATION (BLOCK–HOURS)	4,136	3,426	3,294
COST PER BLOCK–HOUR	343	308	191

SOURCES:

PURCHASE PRICES –	DC8–61 TRANS INTERNATIONAL AIRLINE CORPORATION, STOCK PROSPECTUS, MAY 8, 1967, P 6. 727–200. WOOLSEY [34]. DC9–30 WOOLSEY [34]
UTILIZATION RATES –	CAB, AIRCRAFT OPERATING COSTS AND PERFORMANCE REPORT [28].

Direct airline capital costs are estimated in Table 2. In Table 1, capital costs have been added to current costs to get total direct cost per block-hour. In Section 5, these costs are converted to a basis of seat-miles and passenger-miles for 30 important airline routes.

For purposes of later reference, we consider here the question of appropriate seating capacities for the aircraft under consideration. On the DC9-30, PSA uses a configuration of 110 seats per aircraft; on the 727-200, it gets 158 seats into a plane.[19] Since no intrastate carriers operate a DC8-61, there is no ready source of data on "competitive" seating. However, to achieve an equivalent density (34-inch pitch and 19-inch seat width), and also for a small galley necessary on the long hauls for which the DC8 is used, a capacity of 251 seats would seem appropriate.[20] These seating densities will be used when all airline costs are converted to a passenger-mile and seat-mile basis in Section 5. First, we turn to the problem of estimating indirect airline costs.

---

[18] Evidence on Continental's maintenance policy is presented in Gordon [8], p 89 See Table 1 for the sources for direct operating cost data.

[19] PSA seating capacities are given in Pacific Southwest Airlines Stock Prospectus dated June 11, 1968, p. 7.

4. Indirect airline costs

■ Indirect operating costs include expenses for sales, reservations, aircraft cleaning and fueling, airport rentals, cabin service, and administration. For these expenses, more than one output variable is involved and statistical analysis is necessary to separate their effects.

☐ **Specification of the indirect cost function.** A rather wide variety of different variables has been used to explain indirect airline costs, but they can practically all be distilled into three basic types of variables: (1) a measure of overall capacity of service provided, such as available ton-miles or available seat-miles; (2) a measure of the actual traffic the airline carries, such as revenue ton-miles or revenue passenger-miles; and (3) a measure of expenses which are fixed with stage length or passenger trip length, varying with departures provided or passengers carried.[21]

Of the other types of variables that could be included, perhaps the most likely candidate would be stations served, under the assumption that there are fixed long-run station costs and, hence, increasing returns to scale in the provision of airline terminal services. However, there is no meaningful evidence that such fixed costs exist. A previous study by this writer has shown that there are long-run constant returns to scale in the provision of airport services over a wide range of outputs.[22] Similarly, cost studies by the CAB indicate that stations served play a very minor role in the explanation of indirect airline costs.[23] Therefore, stations served was not included as a variable.

It would be redundant to include in the regression more than one variable in each group, especially given the likely multicollinearity which would result, so one output variable is included from each group, as shown:

$$\mathrm{TCI} = a_0 + a_1\mathrm{ATM} + a_2\mathrm{RTM} + a_3\mathrm{ATD}, \qquad (1)$$

where TCI is total indirect costs; ATM (available ton-miles) measures the total ton capacity of planes flown times distance; RTM (revenue ton-miles) measures total weight of all passengers, baggage, and freight carried; and ATD (available ton-departures) measures the capacity of all departing aircraft, to account for those terminal expenses which are independent of stage length. The estimated coefficients will later be converted to a basis of passenger-miles and passengers carried, using an appropriate load factor.

---

[20] See Watkins [33].

[21] For several recent examples of statistical analysis of indirect airline costs, see Caves [3], pp. 73–77, Straszheim [26], pp. 254–68, and *Local Service Air Carriers' Unit Costs*, March 1970 edition, pp 6–15

[22] See Keeler [11].

[23] See *Local Service Air Carriers' Unit Costs* [30].

☐ **Estimation of the equation.** Cross section data are available for estimation of the equation, but with a limited number of observations (there are only 11 domestic trunk airlines). Pooling of cross section and time series data would clearly be desirable, but it entails all the problems of cross section estimation, time series estimation, and pooling as well. These must be considered before estimation is possible.

If cross sections are to be used at all, it would be wise to take some account of heteroscedasticity in the disturbance term. Firm sizes vary widely and, in this situation, it is highly unlikely that the variance of the disturbance term is independent of firm size. An appropriate assumption here is that the standard deviation of the disturbance term is proportional to the overall scale of operation of the firm. If this is the case, the appropriate procedure is to divide all variables in the regression by an appropriate measure of the scale of operation of the firms.[24] This is essentially the assumption made by Straszheim when he divides all his cost variables by available seat-miles for estimation.[25] For this analysis, all variables are divided by available ton-miles, yielding the following equation for estimation:

$$\frac{\text{TCI}}{\text{ATM}} = \frac{a_0}{\text{ATM}} + a_1 + a_2\frac{\text{RTM}}{\text{ATM}} + a_3\frac{\text{ATD}}{\text{ATM}}.$$

Another problem of cross section cost estimation is that of firm homogeneity. As we have found, there are substantial differences in firm managements, and regional cost variations should add more firm effects. Methods of dealing with these will be discussed below. There is one problem of firm homogeneity which should be dealt with in the context of the data sample, however. That is the problem of cost allocation between domestic and international operations. These two types of operation could have very different cost characteristics; but all trunks have international operations. The CAB has most airlines segregate their cost and output data for domestic and foreign operations; yet, the cost-allocation procedures used are not specified, and may be arbitrary. For 8 out of 11 domestic trunk lines, however, international traffic is not a very important part of the total. For two airlines, Continental and Northwest Orient, international traffic did present problems, and they were excluded from the data sample.[26]

---

[24] See Malinvaud [15], pp. 254–58.

[25] See Straszheim [26], pp. 254–55.

[26] Continental does a very large international charter business for the military, less than half of its total revenue ton-miles in 1968 were for domestic scheduled routes; the problem is that CAB accounts do not segregate the costs of these operations from Continental's domestic costs. For Northwest, the problem was more complicated. When costs were estimated with pooled data as described below, a separate coefficient for available ton-miles was estimated for each airline in the sample to capture interfirm differences in airline costs (using basically the "dummy" variable technique). For Northwest, this coefficient was negative, implying the extremely implausible result that an extra available ton-mile has a *negative* cost. Northwest was the only airline for which this happened, and it was suspected that this was due to faulty cost allocation between domestic and international operations in the CAB accounts.

Since airline technology and factor prices have been changing over time, it would be appropriate to make the time period covered rather short. The three most recent available years of data (1967–1969) were selected. With quarterly data, this makes 12 observations for each airline.[27]

Use of raw quarterly data will yield inefficient estimates of airline costs. The reason is that seasonal variation in airline output is not accompanied by similar variation in overhead costs; an airline does not adjust reservation facilities, station areas, and management staff to meet seasonal variations in output. Since facilities tend to be overcrowded during peak periods and underutilized during off-days, we might assume that the appropriate costs would be those based on a typical day in between. To solve this problem, a standard seasonal adjustment procedure was used. First, a geometric moving average was constructed for each data series of each airline (since different airlines have different seasonal patterns). Next, a mean seasonal deviation from the moving average was constructed for each quarter. Each observation was then divided by its appropriate estimated seasonal weight.

Another problem presented by the use of time series data for cost functions is changes in costs over time. If there were an appropriate price deflator, we would use it; but no such deflator is available. Over the 1967–1969 period there certainly was price inflation, and offhand we should not expect the cost function to remain constant. However, these cost increases were offset by the automation of a good part of airlines overhead, including computerized reservations and automated baggage handling. Given these two offsetting effects, there are statistical methods for determining whether the regression relationship has shifted meaningfully over time.[28]

First, regression coefficients are computed separately from each of the 12 available cross sections. Then, a regression is run with the pooled data. The sums of the squared residuals are computed first for the 12 separate regressions (call this sum A) and then for the pooled data (call this sum B). The following ratio has the F distribution, under the statistical assumptions of ordinary least squares:

$$F = \frac{(B - A)/44}{A/60}.$$

The F ratio was computed for the pooled and unpooled regressions, and was found to be 0.6347. Such a low value provides support for the hypothesis of homogeneity: it implies that it is impossible to reject the homogeneity hypothesis, even at the 50-percent level.

Another minor problem of data homogeneity is that of airline strikes. There were two minor strikes in 1969: one on American Airlines in March and another on Western in August. The solution used here is to include a "strike dummy" variable equal to one for American in 1969 I and for Western in 1969 III, but zero otherwise. This variable is denoted by x_4, and its coefficient by a_4 in the regression.

---

[27] Cost data come from the *Quarterly Report of Air Carrier Financial Statistics* [32], starting with the issue for the quarter ending March 31, 1967 and ending with the issue for the quarter ending December 31, 1969. Output data come from the *Monthly Report of Airline Operating Statistics* [31] for the same period.

[28] See Chow [4].

In using pooled cross section and time series data, what sort of cost function are we likely to be estimating? As previously stated, cross section analysis is usually employed to identify long-run behavior; time series analysis is used to estimate short-run relationships. For estimation of indirect airline costs, this is not such a problem as it might first seem. It has been found by Caves and others that practically all airline indirect costs are variable over an extremely short period of time.[29] If this is the case, then estimation of overhead airline costs with seasonally adjusted quarterly data should yield something very close to a long-run relationship. In any event, there is a straightforward way of determining the extent to which our estimated function represents a long-run relationship. Past evidence indicates reasonably constant returns to scale in the airline industry in the output range under consideration.[30] Accordingly, a long-run cost relationship should be linear and have no constant term. If the estimated fixed costs in the regression equation (as represented by a_0) are small relative to total costs, then we shall have reason to believe that we have estimated a long-run relationship. On the other hand, if the estimated constant term were large relative to total costs, that would imply large fixed costs, consistent only with a short term relationship.

☐ **Estimation techniques.** For pooled cross section and time series data, ordinary least squares will not necessarily yield the most efficient possible estimates, because it assumes that the error terms over time and over cross sections have the same variance. As Kuh first pointed out in the context of investment relationships, there is evidence that the intertemporal error term is quite different in variance from the interfirm error; hence, ordinary least squares yields inefficient estimates.[31]

The most common method of dealing with the problem of firm effects is to include dummy variables in the regression and allow for a different constant term for each firm. A more sophisticated technique, developed by Balestra and Nerlove, treats the firm effect as stochastic and includes it as a part of the error term.[32] Unfortunately, little is known theoretically about the relative desirability of either estimator; since there is no prior reason to prefer one over the other, both techniques are used.[33]

In the equation with differing constant terms, however, if fixed costs are relatively unimportant, it would seem inappropriate to allow for differing values of a_0 among airlines. A more appropriate procedure would be to select one output variable thought to summarize best the scale of operation of the airline, and allow its coefficient to vary from firm to firm. Available ton-miles is just such a variable; since the variables are divided by ATM to correct for heteroscedasticity, the constant in the estimated regression will actually be the coefficient for ATM. Therefore, the coefficient a_1 is allowed to vary from airline to airline.

---

[29] See [3], pp. 79–82. Though airlines may not adjust costs over a period so short as a season, this will cause no problem since our data are seasonally adjusted.

[30] *Ibid.*, pp. 73–83

[31] See [14], pp. 197–214.

[32] There is not sufficient space here to discuss the details of the Balestra-Nerlove estimation procedure and its relationship to the dummy variables technique. For a complete discussion, see [1] and [20].

[33] Balestra and Nerlove suggest several ways of estimating their model. The method used here is the two-stage, best asymptotically normal procedure described in [1] The computer program came from the Regression Analysis Package for Economists, written by W Raduchel.

The regression results are shown in Table 3. It will be noted that the coefficients for the dummy variable model are all significant at the 10-percent level, whereas the same cannot be said of the Balestra-Nerlove equation. The overall reliability of the former method appears greater in this case, if the standard error estimates are to be believed. The R^2 falls precipitously for the Balestra-Nerlove equation, although this is not necessarily significant: the Balestra-Nerlove estimator requires transformation of the dependent variable relative to the earlier procedure, rendering the two R^2's incomparable. One possible reason why the standard errors should be lower in this case with the dummy specification is that interfirm cost variations are important enough here that we gain by taking explicit account of them, rather than putting them into the error term.

TABLE 3

REGRESSION RESULTS FOR INDIRECT AIRLINE COSTS

	COEFFICIENT				
	a_0	a_1	a_2	a_3	a_4
BALESTRA–NERLOVE ESTIMATOR					
ESTIMATE	$151	$0.0803	$0 0369	$1.666	–$0.0282
STANDARD ERROR	662	0.0141	0 0322	0.628	0.0337
R^2 = 0 1068 F(4,103) = 3.0778					
DUMMY VARIABLES TECHNIQUE					
ESTIMATE	$1,721,000	(SEE BELOW)	$0 1414	$1 630	$0.0316
STANDARD ERROR	482,000		0.0102	0.935	0 0049
R^2 = 0 8582 F(4,95) = 143 7					

ESTIMATES OF a_1 FOR EACH AIRLINE·

AIRLINE	ESTIMATE	STANDARD ERROR
AMERICAN	$0 01941	0 00387
EASTERN	0.04391	0.00433
TRANS WORLD AIRLINES	0 02173	0.00411
UNITED	0 02123	0.00444
BRANIFF	0.03231	0.00427
DELTA	0.02025	0 00450
NATIONAL	0.01792	0.00378
NORTHEAST	0.04252	0.00493
WESTERN	0.02874	0.00427

The equation with the dummy variables has yielded results which are both plausible and significant, and also yields interesting results on interfirm cost differences. The firm with the lowest estimate of a_1, National, is headquartered in the South and has a disproportionate number of routes there. Conversely, the firms with the highest coefficients, Eastern and Northeast, have a disproportionate number of routes in the Northeast. Since these are overhead costs, a large part go to wages for clerical workers and ground crews, and we should expect interregional cost differences to affect the results, as evidently they have.

Of the costs estimated for different firms, perhaps the most surprising results are for American Airlines. After National, it has the lowest coefficient (of course, it must be remembered that for reasons already mentioned, two low-cost firms have been excluded from the sample). In a 1962 study of the relative efficiency of various trunk airlines, R. J. Gordon finds American to be consistently the poorest performer, and its improved performance here is difficult to explain. However, Gordon also quotes American's president in the early 1960s as being very much concerned with reducing American's costs.[34] Perhaps he succeeded.

It is important to note that the coefficient a_0 (for the term 1/ATM) is positive and significant, but its estimated value is only 3.7 percent of quarterly overhead costs for the average domestic trunk airline. Hence, these fixed costs are, for practical purposes, negligible. This is strong evidence that we have estimated a long-run relationship, as previously predicted.

Which coefficient for a_1 should be used? To find long-run marginal costs, we should presumably select the minimum-cost airline. But the minimum-cost firm is headquartered in the South and has a larger than normal number of routes there, as well. American Airlines, next in line, has nationwide routes, and there is no reason to believe that its cost should be atypical in any sense. Thus the selected "efficient" marginal cost per available ton-mile is that for American Airlines. The next step is to convert the estimated costs from ton units into passenger units. As of 1968, an average passenger and his baggage weighed 0.0964 tons.[35] To convert cost per revenue ton-mile to cost per passenger-mile, the cost coefficient for revenue ton-miles should be multiplied by this number. To convert available ton-miles and available ton-departures to seat-miles and seat-departures, it is necessary to compare ton capacity and seat capacity for the aircraft to be used in the cost calculations. For the Douglas DC8-61, the Boeing 727-200, and the Douglas DC9-30, a high-density coach seat is equivalent to about 0.108 tons of plane capacity.[36] Using these conversion figures, we find an overhead cost of 17.6 cents per available seat-departure, 1.36 cents per passenger-mile, and 0.226 cents per available seat-mile.

So far, all attention has been focused on overhead cash expenses. Airlines also must invest in overhead capital for their operations, and it is necessary to account for these costs. Doing so presents problems, however, since airline property account data are not separated between domestic and foreign operations. There was only one airline, National, with strictly domestic operations over the 1967–1969 period. Since these capital costs are not a large part of the total, the results should be insensitive to the output variable assigned, and it should not be a cause for concern that there are insufficient degrees of freedom for statistical analysis. National's costs are used for this analysis. Purvis allocates these costs on the basis of plane-hours, and with good reason.[37] A large part of overhead capital costs goes to maintenance of aircraft and training of crews, as well as to spare aircraft parts. All of these costs should vary with plane-hours. but it seems to this writer that large planes should be weighted more heavily than small ones, since they require more servicing and larger crews. Therefore, plane-hours are weighted by plane capacity, and the output variable is available ton-hours (ATH), computed by the following formula:

---

[34] [8], p. 91.

[35] Computed by dividing total revenue ton-miles of passengers and their baggage by total revenue passenger-miles. Data come from the *Handbook of Airline Statistics* [29], 1969 edition, p 108.

[36] More specifically, the numbers are 0.109 for the DC8-61, 0.106 for the 727-200, and 0.111 for the DC9-30 These were computed by dividing the average capacities for these planes by the seating capacities (see the references in notes 20 and 21 above). Ton capacities come from *Aircraft Operating Costs and Performance Report* [28], pp. 16–18. It was felt that these numbers were alike enough to warrant using their mean, 0.108, for these indirect cost calculations.

$$\text{ATH} = (\text{ATM})\,(\text{PH})/\text{PM},$$

where PH is plane-hours and PM is plane-miles. To get a unit cost for overhead capital, we divide National's 1968 gross investment in overhead capital by available ton-hours. The results indicate a gross investment of $12.30 per available ton-hour.[38] With a 12-percent interest rate and 12-year equipment life, this is an annual cost of 0.1576 × $12.30 = $1.96 per available ton-hour.

Finally, allowance must be made for the costs of airline working capital (current assets minus current liabilities). For all domestic U. S. certified air carriers, working capital totaled $422 million.[39] This is equivalent to an investment of 1.13 cents per available ton-mile, or, at a 12-percent interest rate and no depreciation, 0.136 cents per available ton-mile. Assuming 0.108 ton-miles per available seat-mile, we get a working capital cost of 0.015 cents per available seat-mile.

This completes the analysis of indirect airline costs, which are summarized and compiled with direct costs in the next section.

5. Estimates of long-run marginal airline costs

■ In this section, costs are converted to a basis of seat-miles and passenger-miles, and estimates of efficient fares are given.

The first step in compiling and converting airline costs is to convert costs varying with plane-hours to costs per seat-mile. To do this, it is necessary to know block times achievable between the relevant city pairs.

To get an estimate of block time between city pairs, Purvis simply used scheduled times. Although scheduled times do seem to be the best source for estimates of block times, use of them directly will probably yield a downward-biased estimate of costs. A plane is "on time" by CAB definitions as long as it touches the ground within 15 minutes of scheduled time, and on-time ratings for most airlines vary between 75 percent and 90 percent. Block time, however, is measured from the time the plane leaves the gate at one airport to the time it parks at its destination. Hence, using Purvis' method, an airline could have a perfect "on-time" record, yet still yield downward-biased estimates of costs.

---

[37] [23], p. 62

[38] National, in 1968, had 2.77 million available ton-hours and a total gross investment of $34.08 million in ground facilities, equipment, and spare parts Computed from the *Handbook of Airline Statistics* [29], pp. 222, 314.

[39] *Ibid.*, p 107.

To find out how much extra time should be allowed for block time over scheduled time it is necessary to add up all scheduled hours from a timetable and compare them with hours actually traveled. From this procedure, it is possible to determine how long an average trip actually takes relative to its scheduled time. Because of complicated route structures and frequent schedule changes, it is difficult to get annual estimates of scheduled hours. There is one case, however, for which such calculations are relatively easy. Up to 1965, PSA had an extremely simple route structure, and Purvis has managed to calculate an average systemwide scheduled block speed from timetables. Over the 1961–1964 period, it stayed at about 343 miles per hour.[40] We must now estimate PSA's achieved block speed for 1964. This is not easy, since PSA operating data are skimpy. It is possible, however, to make the estimate using the following formula:

$$S_r = \frac{\text{PM}}{(u_a)(n)(b)},$$

where S_r is realized block speed, u_a is airborne aircraft utilization rate, n is the number of planes in PSA's fleet in 1964, b is the ratio of block to airborne hours, and PM is plane-miles. Data are available for PSA for all figures except b, which must be guessed. The value assumed for b is 1.165.[41] The formula yields a realized block speed of 293 miles per hour in 1964.[42] For that year, the average PSA flight was 278 miles long, so the average flight had a scheduled time of 278/343 = 0.81 hours = 49 minutes.[43] But it actually took 278/293 = 0.95 hours = 57 minutes. In other words, the average PSA flight took 8 minutes longer than it was scheduled to take—probably a good record relative to other carriers.[44] We assume for calculations here that the average flight took 10 minutes longer than it was scheduled to take.

Now we must tackle the question of what timetable time to use for each route. Some airlines seem to schedule themselves at higher speeds than others. Furthermore, peak-offpeak variations in congestion make schedules longer during rush hours. Finally, on east-west routes, winds play a part in flying time, making east-bound hauls faster than westbound ones. The solution selected here is to take the average of the fastest and slowest scheduled speeds on each route.

The only run on which this averaging procedure is not used is the Los Angeles-San Francisco run, where we know that PSA has managed, at least in the past, to meet its exceptionally fast schedule

---

[40] Purvis [23], p 37

[41] This figure is the average block to airborne utilization ratio for Continental Airlines DC-9s in 1968. See *Aircraft Operating Cost and Performance Report* [28], p. 55. Continental's DC-9s were selected, first because of Continental's reputation for efficiency, similar to PSA's, and second because they, like PSA's Electras in 1964, were used on short hauls in uncongested airports.

[42] In 1964, PSA's Lockheed Electras had an airborne utilization rate of 8.7 hours per day, and PSA had six such aircraft at the time (Jordan [10], p. 198). PSA operated 6.589 million plane-miles that year (Purvis [23], p. 99). Thus, the average block speed for PSA is

$$\frac{6{,}589{,}000}{(6)(8.7)(366)(1.165)} \approx 293.$$

[43] Purvis [23], p. 32.

[44] *Ibid.*, p. 28.

(the 1968 speed of 45 minutes with jets is 5–10 minutes faster than scheduled speeds of other airlines on the same run) with a very good on-time record. Hence, PSA's time is used for this run, but the 10-minute margin is added. As Jordan and Purvis have implied, these exceptionally fast speeds may be due to special circumstances of PSA, so we can make no assumption that other airlines could achieve equal speeds on similar runs elsewhere. (In fact, increased congestion on the Los Angeles-San Francisco run has forced PSA to increase scheduled running time to 55 minutes today.)[45]

Before computing the fares for a hypothetical "competitive" airline, it is necessary to know a feasible but efficient load factor. As Jordan and Purvis have argued, an airline has substantial latitude available to it in scheduling, so that the load factor it achieves is likely to be the result of a conscious decision on the part of the management.[46] Table 4 shows that the California intrastate carriers managed to keep load factors consistently higher than the domestic trunk carriers over the 1956–1965 period. As Jordan points out, this was the result of careful scheduling on the part of the managements of the airlines involved. In fact, the management of PSA appears to have been very careful to schedule extra flights during periods when traffic volume is high, such as Friday and Sunday nights. But even for the trunk carriers, systemwide load factor data are misleading with reference to high-density routes.

TABLE 4

LOAD FACTORS – CALIFORNIA INTRASTATE VS TRUNK

YEAR	TRUNK AIRLINE LOAD FACTOR (PERCENT)	CALIFORNIA INTRA–STATE LOAD FACTOR (PERCENT)
1951	69.6	69.0
1952	67.1	65.9
1953	64.7	67.1
1954	63.4	69.2
1955	64.1	72.2
1956	64.1	75.7
1957	61.5	80.6
1958	60.0	72.4
1959	61.4	71.1
1960	59.5	71.1
1961	56.2	72.1
1962	53.3	75.3
1963	53.8	72.8
1964	55.4	74.9
1965	55.2	63.3
1966	58.5	
1967	57.2	
1968	53.0	
1969	50.3	
1970	49.3	

SOURCES

TRUNK DATA – 1951–1968: CAB, HANDBOOK OF AIRLINE STATISTICS [29], P. 26.
– 1969–1970: STANDARD AND POOR'S INDUSTRIAL SURVEYS: AIR TRANSPORT [24], MAY 16, 1971 EDITION, P. 70.
CALIFORNIA INTRASTATE DATA – JORDAN [10], P 202.

---

[45] All scheduled speeds have been taken from the *Official Airline Guide* [5]; for 1968 speeds, the January 1969 edition was used.

[46] See Jordan [10], pp. 200–209 and Purvis [23], pp. 86–87.

Caves has shown that load factors tend to be somewhat higher on high-density routes than on low-density routes.[47] As of 1967, the average load factor for flights serving LaGuardia Airport in New York was 59.2 percent, compared with a trunk average load factor of 57.2 percent.[48]

For calculations here, we assume a load factor of 60 percent. This appears to be quite feasible for the trunks on high-density routes, and the California intrastate carriers over the 1956–1965 period maintained one consistently higher. It seems fairly clear that airlines should be able to do better than this, however. It represents a minimum feasible load factor for high-density routes.

In Table 5, all costs are compiled and converted to terms of passengers (passenger-miles, -minutes, or -departures), using the seating densities assumed in Section 4 and the 60-percent load factor. Since we are allowing passenger service costs to vary with passenger-miles, we know that the costs include an allowance for

TABLE 5

COMPILATION OF AIRLINE COSTS

COSTS VARYING WITH BLOCK HOURS	AIRCRAFT TYPE		
	DC8–61	727–200	DC9–30
DIRECT COSTS (DOLLARS)			
CURRENT	579	461	349
CAPITAL	343	308	191
INDIRECT CAPITAL (DOLLARS)			
COST PER AVAILABLE TON–HOUR	1.94	1.94	1.94
AVERAGE TON CAPACITY	28 10	16 10	11.70
TOTAL INDIRECT CAPITAL	55 00	32 00	23 00
TOTAL COSTS VARYING WITH BLOCK HOURS (DOLLARS)	977	801	563
SEATING CAPACITY	251	158	110
COST PER SEAT–MINUTE (CENTS)	6.40	8.45	8.54
COST PER PASSENGER–MINUTE (CENTS) 60 % LOAD FACTOR	10.7	14.1	10 7

COSTS VARYING WITH PASSENGERS OR PASSENGER–MILES	AT FULL LOAD FACTOR	AT 60 % LOAD FACTOR
COST FIXED WITH TRIP LENGTH (CENTS)	17.6	29 3
COSTS VARYING WITH DISTANCE (CENTS)		
COSTS DEPENDENT ON AVAILABLE SEAT MILES	0 226	0 377
COSTS DEPENDENT ON PASSENGER MILES	1.36	1.36
WORKING CAPITAL COST	0 015	0.025

TOTAL MILEAGE–RELATED COST AT 60–PERCENT LOAD FACTOR
C_m = 29 3 ¢ plus 1.76 ¢ times m, WHERE C_m = MILEAGE–RELATED COST PER PASSENGER AND m = DISTANCE OF TRIP IN MILES

---

[47] [3], pp. 77–78.

[48] See Carlin and Park [2]

meal service on long hauls, but not on short hauls. Accordingly, the DC8-61 seating configuration allows for galley space, whereas the configurations for the shorter-haul planes do not. The DC9s were designed for short hauls, the 727s for medium hauls, and the DC8s for long hauls. For flights under 350 miles, DC9s were used in the calculations; for the 350–900 mile range, 727–200's were used; and for flights over this distance, the calculations assume use of a DC8-61.

In Table 6, hypothetical unregulated coach fares are calculated and shown for the 30 largest airline routes in continental United States, as of 1968. Also, a similar fare is calculated for the Los Angeles-San Diego market, since PSA's fares in this market will give a check on the calculations. Alongside the unregulated fare estimates are shown the actual coach fares. No effort is made here to adjust for interregional cost differences, except to allow for airport and airways congestion, which is probably the most important source of such differences.

The estimates are extremely close to the actual fares. For the Los Angeles-San Francisco route, the estimated fare is $13.87, compared with an actual 1968 fare of $13.50. For the Los Angeles-San Diego route, the estimated fare is $6.50, compared with an actual fare of $6.35. In each case, the model comes within 3 percent of predicting actual fares. The cost model developed and estimated here does a good job of predicting unregulated fares on short hauls. Unfortunately, there are no unregulated long-haul markets with which to compare results. However, in April 1967, World Airways, a supplemental charter carrier, did *attempt* to enter the transcontinental market. In an application to the CAB, it proposed jet service between California and the East Coast with a fare of $79.50.[49] This compares with estimated unregulated fares of $79.20 on the New York-Los Angeles route and $82.08 on the New York-San Francisco run. Again, the prediction error is small; thus the model has done a good job of predicting unregulated fares on runs ranging in length from 109 miles to 2,500 miles. It should give reasonably good estimates of what unregulated fares would be in other high-density markets.

Also in Table 6, the markup of existing coach fares over unregulated fares is shown, in percentage terms. There are two striking things about these markups. The first is their size—none are below 20 percent for regulated markets, and they reach a staggering 94.5 percent on long-haul routes, such as New York-Miami. Second, this markup is not random relative to trip length—there is a pronounced tendency for markups to rise with distance.

These calculations are based on standard fares, and to some extent, this is misleading. Many nonbusiness travelers have been able to take advantage of excursion fares which, in 1968, were applicable to nearly all trunk routes and which allowed a 25-percent discount from standard fares. Under these circumstances, any route with a markup below 33 percent will offer service at a loss (relative to measured long-run marginal costs) with excursion fares. Note that this includes many of the short hauls in Table 6. Little wonder that excursion fares have since been eliminated for

---

[49] See Jordan [10], p. 113.

TABLE 6

AIRLINE COSTS AND FARES FOR 30 CITY PAIRS (1968 PRICES)

CITY PAIR	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
BOSTON–NEW YORK	63	DC9	190	$ 9.01	$ 3.64	$12.65	$16	26.5
NEW YORK–WASHINGTON	70	DC9	216	10.01	4.09	14.10	17	20.6
CHICAGO–NEW YORK	130	727	721	18.33	12.98	31.31	44	40.5
MIAMI–NEW YORK	160	DC8	1093	17.12	19.54	36.66	71	94.5
LOS ANGELES–SAN FRANCISCO	55	727	347	7.76	6.11	13.87	13.50	–2.7
LOS ANGELES–NEW YORK	334	DC8	2453	35.74	43.46	79.20	145	83.1
LAS VEGAS–LOS ANGELES	55	DC9	226	7.87	4.27	12.14	15	23.6
DETROIT–NEW YORK	96	727	489	13.54	8.90	22.44	32	42.6
NEW YORK–SAN FRANCISCO	341	DC8	2574	36.49	45.59	82.08	145	76.7
NEW YORK–PITTSBURGH	77	DC9	330	11.01	6.40	17.41	24	37.8
CLEVELAND–NEW YORK	88	727	410	12.40	7.51	19.91	28	40.6
CHICAGO–LOS ANGELES	239	DC8	1740	25.57	30.92	56.49	90	59.3
BUFFALO–NEW YORK	83	DC9	289	11.70	5.38	17.08	22	28.7
BOSTON–WASHINGTON	90	727	406	12.69	7.44	20.13	27	34.1
CHICAGO–DETROIT	68	DC9	238	9.72	4.48	14.20	19	33.8
CHICAGO–MIAMI	175	DC8	1188	18.07	21.20	39.27	74	88.3
CHICAGO–MINNEAPOLIS	80	727	345	11.28	6.36	17.64	26	47.4
SAN FRANCISCO–SEATTLE	94	727	672	13.25	12.12	25.37	43	69.5
CHICAGO–ST. LOUIS	73	DC9	256	10.44	4.80	15.24	20	31.2
CHICAGO–PHILADELPHIA	118	727	669	16.64	12.07	28.71	41	42.8
LOS ANGELES–SEATTLE	145	DC8	959	15.52	17.17	32.69	61	86.6
ATLANTA–NEW YORK	130	727	755	18.35	13.58	31.93	50	56.8
BOSTON–PHILADELPHIA	70	DC9	280	10.01	5.22	15.23	21	37.8
NEW YORK–ROCHESTER	70	DC9	252	10.01	4.73	14.74	21	42.5
CHICAGO–WASHINGTON	117	727	591	16.50	10.69	27.19	50	83.9
CHICAGO–SAN FRANCISCO	249	DC8	1854	26.64	32.92	59.16	105	76.3
CHICAGO–CLEVELAND	79	DC9	311	11.30	5.76	17.06	22	29.0
DALLAS–HOUSTON	63	DC9	240	9.01	4.52	13.53	19	40.4
NEW YORK–SYRACUSE	64	DC9	197	9.15	3.76	12.91	16	23.9
BOSTON–CHICAGO	143	727	860	20.16	15.43	35.59	50	40.5
LOS ANGELES–SAN DIEGO	30	DC9	109	4.29	2.21	6.50	6.35	–2.3

COLUMN DEFINITIONS

(1) TYPICAL SCHEDULED TIME plus 10 MINUTES (SEE SECTION 5) WHEN PEAKLOAD FARES ARE AVAILABLE, COSTS COMPUTED ARE BASED ON TYPICAL PEAK TIME AND FARE USED FOR COMPARISON FOR PEAK PERIOD

(2) PLANE TYPE ASSUMED DC8 = DOUGLAS DC8–61, 727 = BOEING 727–200, DC9 = DOUGLAS DC9–30.

(3) AIRLINE DISTANCE (MILES)

(4) COSTS VARYING WITH TIME COLUMN (1) times COST PER PASSENGER MINUTE FROM TABLE 7

(5) COSTS VARYING WITH DISTANCE: COMPUTED WITH MILEAGE FROM COLUMN (3) AND FORMULA AT BOTTOM OF TABLE 7

(6) TOTAL LONG–RUN MARGINAL COST COLUMN (5) plus COLUMN (6)

(7) FARE. LOWEST REGULARLY–AVAILABLE DAYTIME FARE (COACH OR ECONOMY), ONE–WAY

(8) PERCENTAGE MARKUP (COLUMN (7)/COLUMN (6) times 100) –100

SOURCE: TIMETABLE TIMES, FARES, AND DISTANCES COME FROM THE OFFICIAL AIRLINE GUIDE [5], JANUARY, 1969 EDITION.

trips under 1,500 miles; the discount has also been lowered to 15 percent.

Notably, the 60-percent load factor chosen here is lower than that achieved by the California carriers anytime between 1950 and 1965. Yet suppose that unregulated carriers on other routes could not achieve even that load factor. Suppose, furthermore, that on-time performance were more difficult outside California, and hence block times were longer than average scheduled times by more than 10 minutes. How would this affect the resulting cost-based fares? For the sake of argument, assume that our hypothetical unregulated airline managed to fill only half of its seats on the average (more in line with what trunk airlines have actually achieved). Suppose, also, that each flight took 20 minutes longer than it was scheduled to take, rather than the 10 minutes assumed before. Table 7 shows cost-based fares estimated with these assumptions of lower overall efficiency, for the top 25 interstate markets. The results still reveal a positive price-cost margin on most routes, and for long-haul routes, markups are still over 50 percent. This implies that for long-haul fares our results are still not very sensitive to assumptions of load factors or on-time performance.

TABLE 7

AIRLINE COSTS AND PRICES WITH A 50–PERCENT LOAD FACTOR[a] (1968 PRICES)

CITY PAIR	COST	FARE	MARKUP
BOSTON–NEW YORK	$16.33	$16	−2 0
NEW YORK–WASHINGTON	18.00	17	−5.9
CHICAGO–NEW YORK	37.23	44	18.0
MIAMI–NEW YORK	42.22	71	68.2
LOS ANGELES–NEW YORK	89 52	145	62.0
LAS VEGAS–LOS ANGELES	15 63	15	−4.2
DETROIT–NEW YORK	27 26	32	17.4
NEW YORK–SAN FRANCISCO	92.64	145	56.5
NEW YORK–PITTSBURGH	21 30	24	12.7
CLEVELAND–NEW YORK	24.45	28	14.5
CHICAGO–LOS ANGELES	64.24	90	40.0
BUFFALO–NEW YORK	21.57	22	2.0
BOSTON–WASHINGTON	24 72	27	9.2
CHICAGO–DETROIT	18 07	19	5 1
CHICAGO–MIAMI	45.89	74	61.3
CHICAGO–MINNEAPOLIS	21.91	26	18.7
SAN FRANCISCO–SEATTLE	30 29	43	42.0
CHICAGO–ST LOUIS	19 25	20	3 9
CHICAGO–PHILADELPHIA	34 29	41	19.6
LOS ANGELES–SEATTLE	37 84	61	61.2
ATLANTA–NEW YORK	37 90	50	31.9
BOSTON–PHILADELPHIA	19.18	21	9.5
NEW YORK–ROCHESTER	18 67	21	12 5
CHICAGO–WASHINGTON	32.68	50	53.0

[a] COMPUTED AS IN TABLE 6, EXCEPT WITH A 50–(RATHER THAN 60–) PERCENT LOAD FACTOR AND WITH 20 MINUTES OF EXTRA BLOCK TIME OVER SCHEDULED TIME

There is something peculiar about this result, however. Throughout the 1960s the airlines maintained load factors of at least 50 percent, yet they earned only normal profits on the average. That would seem to be inconsistent with the results shown in Table 7. There are several reasons for this discrepancy. First, there is clearly some cross subsidy from long-haul to short-haul service, as indicated in Table 7. Second, computations here are based on use of the most efficient available aircraft for each stage length; obviously, not all aircraft were that efficient. Third, we have assumed high-density coach seating configurations of the sort used by the California intrastate carriers, whereas most interstate flights offer substantial amounts of first-class space, as well. Clearly, a 50-percent load factor with a mixed (coach and first class) configuration entails fewer passengers per plane, and a higher cost per passenger, than does the all-coach configuration assumed here. If first-class fares yielded the same amount of revenue per unit of airplane space as coach fares, the high price-cost margins estimated here, combined with a 50-percent load factor, would imply high profits, as well.[50] As it is, however, first-class fares are not so lucrative as coach fares, and first-class load factors are not so high as in coach. In 1968, for example, a first-class seat took up 1.68 times as much space as a coach seat, yet first-class fares yielded only 1.36 times as much revenue per mile as coach fares.[51] Furthermore, in the same year, the coach load factor was 56 percent, compared with only 43 percent for first class.[52] Overall, the revenue yield of a first-class seat-mile was 2.98 cents, whereas the yield of 1.68 coach seats (one first-class space equivalent) was 4.80 cents per mile. To a large extent, then, it is because of these lower first-class yields that the apparently high price-cost margins derived from a 50-percent load factor in an all-coach configuration do not yield a high profit rate with the same load factor in mixed configurations. In a very real sense, profits from coach service are cross-subsidizing first class.

Using data for 1958, Caves found (on the highest density routes) markups of 56 percent on short to medium hauls (201–400 miles), 67 percent on medium to long hauls (401 to 1,200 miles) and 64 percent on the longest hauls (over 1,200 miles).[53] Since Caves used first-class fares and included no return on capital in his costs, his results are not directly comparable with the ones presented here. However, the ranges of the markups can be compared. They appear to rise much more sharply with distance in 1968 than in 1958. The reason for this greater range of markups in 1968 would appear to stem from relatively more economical long-haul aircraft, more congestion on short-haul, high-density routes, and failure of fares to adjust to these changes.

Since 1968, however, fares have been adjusted substantially, not only upwards, but raising short-haul fares relative to long hauls.[54] How have these changes, plus inflation of factor input prices and technical progress, affected airline markups? One way to answer this question would be to find the percent fare increase since 1968 in the California market, and to apply this same percentage increase to the hypothetical unregulated airline fares in Table 6.[55] This procedure is likely to bias upwards estimates of

---

[50] This statement reflects only what would happen if first class fares were higher *and* if load factors remained at 50 percent. The actual effects of a first-class fare increase depend, of course, on the demand elasticity for first-class service.

[51] This discriminatory pricing is not unique to the year 1968. It goes back to 1960, when the airlines feared that without an increase in jet coach fares relative to first class, businessmen might stop buying first-class tickets. See Caves [3], p 166.

[52] First-class load factors have been similarly lower than coach for over 20 years. See the *Handbook of Airline Statistics* [29], 1969 edition, Tables 22, 25.

[53] [3], p. 409. Caves computes the ratio of cost to price, rather than the markup

[54] For a good summary of all fare changes since the 1920s, see *Standard and Poor's Industrial Surveys: Air Transport* [24], May 16, 1971 edition, pp. A69–71.

[55] The Los Angeles-San Francisco fare went up from $13.50 in 1968 to $15.28 in 1972, for an increase of 13.2 percent. See Tables 6 and 8.

TABLE 8

1972 FARES AND COSTS

ROUTE	1972 FLYING TIME (MINUTES)	1972 COST (DOLLARS)	1972 FARE (DOLLARS)	PERCENT MARKUP
BOSTON–NEW YORK	62	14.31	22 22	55.3
NEW YORK–WASHINGTON	69	15.96	24.07	50.8
CHICAGO–NEW YORK	128	35.66	54.63	53.2
MIAMI–NEW YORK	170	41.76	76.85	84.0
LOS ANGELES–SAN FRANCISCO	65	15.70	15.28	–2.7
LOS ANGELES–NEW YORK	340	90 21	150.93	67.3
LAS VEGAS–LOS ANGELES	58	13 83	25.00	80.8
DETROIT–NEW YORK	105	25.56	41.67	63.0
NEW YORK–SAN FRANCISCO	345	93.49	150 93	61.4
NEW YORK–PITTSBURGH	74	19.83	32.41	63.4
CLEVELAND–NEW YORK	94	22.68	36.11	59.2
CHICAGO–LOS ANGELES	245	64.34	101.85	58.3
BUFFALO–NEW YORK	75	19.45	28.75	47.8
BOSTON–WASHINGTON	90	22 93	35.18	53.4
CHICAGO–DETROIT	69	16 17	25.00	54.6
CHICAGO–MIAMI	180	44 73	82.42	84.3
CHICAGO–MINNEAPOLIS	75	20.09	32.41	61.3
SAN FRANCISCO–SEATTLE	115	28.90	47.22	63 4
CHICAGO–ST LOUIS	63	17.36	25 93	49 6
CHICAGO–PHILADELPHIA	130	32.70	50.45	54.3
LOS ANGELES–SEATTLE	144	37.23	62 04	66 6
ATLANTA–NEW YORK	130	36.37	56.45	55.2
BOSTON–PHILADELPHIA	74	17.35	28.10	62.0
NEW YORK–ROCHESTER	69	16.79	25.93	54.4
CHICAGO–WASHINGTON	118	30.97	47 22	52.5
CHICAGO–SAN FRANCISCO	252	67.38	113.88	69.0
CHICAGO–CLEVELAND	79	19.43	30.56	57.3
DALLAS–HOUSTON	58	15.41	18 52	20.2
NEW YORK–SYRACUSE	65	14.70	23.15	57.5
BOSTON–CHICAGO	140	40.54	62.96	55.3

AVERAGE FLYING TIMES ARE COMPUTED AS BEFORE, AND THE FIGURE SHOWN IS AVERAGE SCHEDULED TIME plus 10 MINUTES 1972 COSTS ARE 1968 COSTS times 1.132 (THE PERCENTAGE INCREASE OF PLANE FARES FROM 1968 TO 1972 ON THE LOS ANGELES–SAN FRANCISCO ROUTE) THE MARKUP IS COMPUTED AS IN TABLE 6

SOURCE: 1972 FLYING TIMES AND FARES CAME FROM THE OFFICIAL AIRLINE GUIDE [5], FEBRUARY, 1972 EDITION.

interstate airline costs as of 1972: PSA's costs are likely to have gone up more than those of most other airlines because of increased congestion on the California routes. Of course, the east coast routes are still more congested than those in California, but while Northeast Corridor and other flying times have remained more or less constant since 1968, PSA's scheduled flying time between Los Angeles and San Francisco has risen from 45 minutes in 1968 to 55 minutes in 1972 (to see this, compare the flying times shown in Tables 6 and 8). The results of the calculations are shown in Table 8.

According to Table 8, the situation has hardly improved regarding the size of interstate airline markups, which now range from 45 percent on short and medium hauls to 84 percent on long hauls. Hence, the main effect of the recent fare increase has been to raise markup on short hauls to at least 45 percent without substantially affecting the already high markup on long hauls.

The one exception to this dreary story is the Dallas-Houston route which, of course, is intrastate. Southwest Airlines has for some time sought the legal right to operate on this and other Texas intrastate routes. In 1971, it finally won this right and, when it entered the market, it immediately lowered the fare from $25 to $18.52.[56] Although this is still a 20-percent markup over the hypothetical unregulated fare estimated here for Dallas-Houston, it represents an important improvement. In this case, the cost model may be giving a somewhat downward-biased prediction of unregulated fares. But, on the other hand, it could well be that the price-cutting in Texas is not over yet.

6. Conclusions and policy implications

■ The results of this study show that regulation has indeed had a strong effect on airline fares. In terms of the actual quantitative effects of regulation, our results fall between those of Jordan and Purvis.

Jordan, using data for 1965 (when fares were basically the same as in 1968), finds a markup of 47 to 89 percent of regulated over unregulated fares. Purvis, using data for 1964, finds that Northeast Corridor fares would be little if any lower than California-type fares with unregulated operation. The results of our study indicate that as of 1968, the markup was as low as 20 percent on short-haul regulated routes and 95 percent on long-haul regulated routes. For 1972, on the other hand, the results are much closer to those of Jordan for the mid-1960s—a markup of 45 to 84 percent. Our results differ rather strongly from those of Caves, since he implies that deregulation would probably bring about little in the way of improvement in airline market performance.

Many past investigations of market performance (including Caves') have taken low accounting profits as a sign of good performance. This study shows how deceptive accounting profits can be as indicators of market efficiency.

Over most of the years of its existence, Pacific Southwest Airlines has generally been more profitable than a typical trunk airline, but it has also charged lower fares.[57] As previously stated, the trunk airlines have, over the years, earned an approximately normal profit. These high costs seem to stem mostly from excess capacity. The interstate trunk carriers have consistently maintained a load factor substantially below that on the unregulated California intrastate routes. It would appear that with fares set at high, cartel levels, the airlines have competed away profits through excess capacity. Since more frequent flights mean more business, there is an incentive to compete through flight frequency instead of prices, and the resulting excess capacity appears to keep profits down to a normal level. Furthermore, relatively high price-cost margins for air coach service do not translate into high overall profits, because first-class fare yields are substantially lower, and if costs are allocated between coach and first-class service on the basis of airliner space, there appears to be some cross-subsidization from coach to first class.[58]

---

[56] That is the fare now on the Dallas-Houston route. See the *Official Airline Guide* [5], February 1, 1972 edition, p. 184.

[57] See Jordan [10], p. 192.

There is also some evidence that the high-density routes studied here are more profitable than the aggregate airline rates of return shown here would indicate, the excess profits being used to cross-subsidize service on less profitable, low-density routes. Indeed, this has been one of the major principles of CAB regulation since 1938. But just how much cross-subsidization goes on any more? Since trunk airlines are often allowed to abandon low-density routes to local-service airlines, it is not clear what incentive they have to keep unprofitable routes now, and they have abandoned such routes in large numbers.[59] Thus though there may be some cross-subsidization, it is not likely to account for a large part of the price-cost discrepancies observed here.

It appears then that airline regulation extracts high costs in inefficiency on high-density routes. What economic reasons are there in favor of continued regulation?[60]

The first reason for regulation, and the one given most often by airline executives, is that in the absence of regulation, cutthroat competition would prevail, putting everyone into bankruptcy. This is allegedly what happened in the 1930s without regulation. As Caves showed ten years ago, however, the near bankruptcy of the airlines in the 1930s was caused mainly by the competitive bidding system used by the Post Office in alloting airmail subsidies (and by the fact that no airline was viable without subsidies in the first place).[61]

A second reason for regulation is that it allows for service to a much larger number of cities than unregulated service would allow, and by doing so it reduces the need to subsidize small-town service with tax revenues. By this argument, regulation is a means of taxation to subsidize unremunerative but still socially desirable services.[62] There are two potential objections to this argument as a reason for continuation of regulation. First, how much cross-subsidization actually occurs as a result of this? It has already been asserted here that there is likely to be little of such cross-subsidization in effect now, despite the efforts of the CAB. Second, is there any need to cross-subsidize low-density service in the first place?

An economic argument for such subsidies would probably have to hinge on the notion that low-density airline service is an indivisible commodity; one can have regularly scheduled, daily airline service or not have it, with nothing in between. Yet the facts indicate that low-density airline service is not really that indivisible. In recent years, commuter or air taxi carriers have managed to provide profitable (unsubsidized) service on many routes on which even the subsidized local-service carriers have not been able to break even.[63] Given that this is the case, it is difficult to see any reason for subsidizing airline service at all.

---

[58] No attempt is made here to pass judgment on the first-class–coach fare differential as a profit-maximizing strategy. That is a topic for further research. It is worth noting, however, that none of the intrastate carriers in Texas or California offers first-class service, whereas trunk carriers competing on the same routes do.

[59] See Caves [3], p. 410.

[60] Only the more important economic reasons for regulation are discussed here. For a more complete discussion of the goals of airline regulation, see Caves [3], pp. 123–30.

[61] *Ibid.*

[62] See Posner [22].

A third reason given for airline regulation is safety. Certainly, new airlines should be required to meet rigid safety standards; the Federal Aviation Administration now requires that of all carriers, interstate and intrastate (and there is no evidence that the existing intrastate carriers are more accident-prone than any others). The question is whether airlines which meet FAA safety standards should be allowed to enter markets and set fares freely. From the viewpoint of safety, it is difficult to see why fares and entry should be controlled.

Stigler, Jordan, and others have argued, finally, that airline regulation is basically a government-enforced cartel existing for the benefit of the regulated firms. If this is so, regulation has not accomplished this either, since, on the average, airlines have earned normal profits for the past 35 years.[64]

Given that airline regulation on high-density routes extracts high social costs and confers very few benefits on anyone, the case against it is strong indeed.

References

1. Balestra, P. and Nerlove, M. "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas." *Econometrica,* Vol. 34, No. 3 (July 1966), pp. 585–612.
2. Carlin, A. and Park, R "Marginal Cost Pricing of Airport Runway Capacity." *The American Economic Review,* Vol. 60, No. 3 (June 1970), pp. 310–19.
3. Caves, R. E. *Air Transport and Its Regulators.* Cambridge, Mass.: Harvard Univ. Press, 1962
4. Chow, G. C. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica,* Vol. 28, No. 3 (July 1960), pp. 591–605.
5. Donnelley, Reuben H. Corp. *Official Airline Guide.* New York: Reuben H. Donnelley Corp , semimonthly.
6. Eads, G., Nerlove, M. and Raduchel, W. "A Long-Run Cost Function for the Local Service Airline Industry." *The Review of Economics and Statistics,* Vol 51, No. 3 (August 1969), pp. 258–70.
7. Feldman, J. M. "Commuters: Where Do They Go From Here?" *Airline Management and Marketing,* Vol. 2, No. 12 (December 1970), pp 14–17
8. Gordon, R. J. "Airline Costs and Managerial Efficiency," in National Bureau of Economic Research, *Transportation Economics* [17], pp. 61–94.
9. Johnston, J. *Statistical Cost Analysis.* New York: McGraw-Hill Book Co., 1960.
10. Jordan, W. A. *Airline Regulation in America.* Baltimore, Md.: Johns Hopkins Univ Press, 1970.
11. Keeler, T. E. "Airport Costs and Congestion." *American Economist,* Vol. 14, No. 1 (Spring 1970), pp. 47–53.
12. ———. "Railroad Cost Functions: An Empirical Study." Unpublished bachelor's thesis. Reed College, 1967.
13. ———. "Resource Allocation in Intercity Passenger Transportation." Unpublished Ph.D. dissertation. Massachusetts Institute of Technology, 1971.
14. Kuh, E. "The Validity of Cross Sectionally Estimated Behavior Equations in Time-Series Applications," *Econometrica,* Vol. 27, No 2 (April 1959), pp. 197–214.

---

[63] See Feldman [7].

[64] With existing excess capacity, deregulation could cause temporary airline losses. But the importance of this problem depends on the demand elasticity for airline service: lower fares could conceivably attract enough new traffic to eliminate the excess capacity.

15. Malinvaud, E. *Statistical Methods of Econometrics.* Chicago: Rand McNally & Co., 1966.
16. Meyer, J. R. and Kraft, G. "The Evaluation of Statistical Costing Techniques as Applied in the Transportation Industry." *The American Economic Review,* Vol. 51, No. 2 (May 1961), pp. 313–34.
17. Meyer, J. R. and Straszheim, M. R. *Techniques of Transport Planning.* Washington, D. C.: The Brookings Institution, 1971.
18. Moody, John. *Moody's Transportation Manual.* New York: John Moody, 1971.
19. National Bureau of Economic Research. *Transportation Economics.* Universities-National Bureau of Economic Research Conference Volume. New York: Columbia Univ. Press, 1965.
20. Nerlove, M. "Problems and Methods in the Estimation of Dynamic Economic Relationships from a Time Series of Cross Sections," in M. Nerlove, *Estimation and Identification of Cobb-Douglas Production Functions,* Chicago: Rand McNally & Co., 1965, pp. 157–90.
21. Peck, M. J. and Meyer, J. R. "A Fair Rate of Return for Regulated Industries," in National Bureau of Economic Research, *Transportation Economics* [17], pp. 199–239.
22. Posner, R. A. "Taxation by Regulation." *The Bell Journal of Economics and Management Science,* Vol. 2, No. 1 (Spring 1971), pp. 22–50.
23. Purvis, G. "A Study of Profitable Short-Haul Air Transportation in Dense Markets." Unpublished masters thesis. Sloan School of Management, Massachusetts Institute of Technology, 1967.
24. Standard and Poor. *Standard and Poor's Industrial Surveys. Air Transport,* New York, Standard and Poor, annual.
25. Stigler, G. J. "The Theory of Economic Regulation." *The Bell Journal of Economics and Management Science,* Vol. 2, No. 1 (Spring 1971), pp. 3–21.
26. Straszheim, M. R. *The International Airline Industry.* Washington, D. C.: The Brookings Institution, 1969.
27. ———. *Techniques of Transport Planning.* Washington, D. C.: The Brookings Institution, 1971.
28 ———. *Aircraft Operating Costs and Performance Report.* Washington, D. C.: U. S. Government Printing Office, Vol. 4 (August 1970).
29. U. S. Civil Aeronautics Board. *Handbook of Airline Statistics.* Washington, D. C.: U. S. Government Printing Office, annual.
30. ———. *Local Service Air Carriers' Unit Costs.* Washington, D. C.: U. S Government Printing Office, semiannual.
31. ———. *Monthly Report of Airline Operating Statistics.* Washington, D. C.: U. S. Government Printing Office.
32. ———. *Quarterly Report of Air Carrier Financial Statistics.* Washington, D. C.: U. S. Government Printing Office.
33. Watkins, H. D. "The Stretched Versions of DC 8 Planned." *Aviation Week,* Vol. 28 (April 12, 1965), pp. 54–67.
34. Woolsey, J. P. "Boeing 727-200: $1.5 Billion Success Story." *Aviation Week,* Vol. 32 (February 3, 1969), pp. 42–43.

Quality variation when prices are regulated

Lawrence J. White

Assistant Professor of Economics
Princeton University

Quality has been a neglected area of public utility regulation analysis. Many authors are uneasy about the quality levels that result from regulation but they frequently can only conclude that "it is uncertain whether consumers found this quality to be worth the price that they paid."

This paper includes a quality variable in a simple model of competition and monopoly under price regulation. Competitors compete away any potential revenue above the cost of the basic output by offering higher quality; a monopolist chooses the profit maximizing quality level. Some important conclusions are as follows: (1) A regulated competitive industry will offer more quality per unit of basic output than will an equally regulated monopolist; (2) the competitor's quality offerings will vary directly with the regulated price; (3) regulation induces a uniformity of quality offerings, whereas an unregulated industry would offer a variety; and (4) virtually all consumers must be worse off when a competitive industry is regulated, because they lose quality choice. If a separate unregulated price can be charged for quality, results comparable to an unregulated industry may be achieved.

1. Introduction

■ This paper will deal with the following question: What happens to quality variables of output when the price of output is fixed by external regulation? This question has been partially treated by Archibald and by Stigler.[1] Archibald, in particular, despairs of finding any predictive content in models containing quality variables.

Section 2 of this paper will discuss the general nature of quality variables. Sections 3 and 4 will analyze a specific regulatory model, in which the price of the basic output is regulated, and it is the only price that can be charged regardless of quality. Section 5 will examine a second regulatory model, in which the price of the basic output is regulated and extra (nonregulated) prices can be charged for quality.

Lawrence J. White received the A.B. degree from Harvard (1964) and the M.Sc. (Econ.) from the London School of Economics (1965). He returned to Harvard and received the Ph.D. degree (1969). Dr. White has participated in the Harvard Development Advisory Service as an advisor to the governments of Pakistan and Indonesia.

An earlier version of this paper was presented at the AT&T Conference on Public Utility Regulation, Dartmouth College, September 1971. The author is indebted to his colleagues at Princeton and to Professor Paul MacAvoy for their helpful comments and suggestions.

[1] See [3] and [6].

2. Definition of quality variables

■ Quality variables are usually treated as inputs into the production process. However, we can think of a quality variable as essentially a second *output* of the firm. If the firm's primary output is "air transportation," the second output would be "inflight meals" or "inflight movies." If the basic output is "90-octane gasoline," the second output is "extra octanes." If the firm's basic output is "sale of clothes in an austere showroom," the second output is "extra comfort and ambiance while selling." Many other examples are possible.

As a second output of the firm, a quality variable has two special characteristics. First, it influences the demand for the firm's primary output. It may influence the overall market demand for the output, or at the limit it may simply influence the division of an unchanged market demand among firms. In turn, the quality variable may have a demand function that is influenced by the level of the firm's primary output. Second, it is tied to the primary output, both in production and in consumption.

Further, let us assume that consumers view this second output as identical among firms in an industry. This means that all airline meals which cost the same are seen to be identical, as are all gasoline octanes, etc. The only possible difference among firms is the *number* of meals or octanes offered (the level of quality), which is important to consumers.

The principles of the model to be discussed in this paper can be generalized to any industry. However, for expositional ease, I shall call the industry under discussion "the airline industry," whose primary output is measured by the number of passengers transported between a given pair of cities.

Two kinds of quality variables can accompany the primary output. First, there are "meals," a quality variable that is produced and consumed in units that are tied directly to each unit of primary output. The producer sets the quality level by deciding how many meals per passenger he will provide. The total number of meals will then vary directly with the total number of passengers. Similar quality variables would be octanes (if the primary output were gasoline) or chrome (for automobiles). The second type of quality variable is "advertising." It is not tied to units of primary output but instead is set at an overall level. In the short run it does not vary directly with primary output. Similar quality variables would be downtown ticket offices or luxurious showrooms.

The difference between these two types of variables is analogous to the difference between Marshallian variable costs and fixed costs. In the long run, all costs are variable and all quality variables behave like meals. But in the short run some costs are fixed and some quality variables do not vary directly with output.

3. The "meals" model

■ Suppose an industry composed of n firms produces a basic output, transportation, and a second output, meals, which is a quality aspect of the primary output. The overall market demand functions for transportation and meals are likely to be complementary; at the limit they will be independent. They can be stated in the following forms:

$$Q_T = Q_T(P_T, P_M) \tag{1}$$

$$Q_M = Q_M(P_T, P_M), \tag{2}$$

where Q_T and P_T are the quantity and price of transportation, Q_M is the quantity of meals per passenger per trip, and P_M is the price of a meal. The total quantity of meals demanded, of course, will be $Q_M \cdot Q_T$.

Further, it is assumed that the unit costs of transportation and meals, C_T and C_M, are constant over all relevant ranges of output. These costs include a "normal" profit on invested capital. Accordingly, the profits, π, of one perfect competitor among n of identical size in the industry can be stated as

$$\pi = P_T q_T + P_M q_M q_T - C_T q_T - C_M q_M q_T, \tag{3}$$

where q_T and q_M are the outputs of the individual firm. The profit maximizing results are familiar: a perfectly competitive industry would set the prices of basic transportation and of meals equal to their unit costs, and the corresponding quantities would be sold; a monopolist ($n = 1$) would set a higher price for at least one good, if not both, and sell a smaller quantity of one or both.[2]

A second important result emerges from this model. In the absence of problems of indivisibilities or economies of scale, the competitive industry would offer a wide choice of quality levels.[3] Flights with 1, 2, 3, or m meals served per passenger would be offered as long as passengers were willing to pay the price charged for the meals. Effectively, flights with more meals would carry a higher total price, flights with fewer meals a lower total price. Or, if passengers on a given flight did not all have to consume the same number of meals, airlines would charge a basic price for the basic transportation and sell meal tickets at the door. Each passenger could buy as many or as few meal tickets as he wished. The monopolist might not be willing to offer a comparable variety of quality levels because of perceived cross-elasticities and joint profitability effects. He would refrain from offering some quality levels if the purchase of these transportation-and-meals combinations would add less to profits than would the other combinations that consumers would buy if the former combinations were not available.[4]

Now, suppose the industry comes under price regulation, and the regulating authority sets a single price, P_T^*, for "transportation." Firms are free to provide any level of quality but they can only charge P_T^*, regardless of quality. Finally, the regulating authority forbids new entry into the industry.

Under these circumstances, we have to change our transportation demand function slightly:

$$Q_T = Q_T(P_T^*, Q_M). \tag{1a}$$

---

[2] Because of the complementary nature of the goods, a monopolist might wish to sell one as a "loss leader." See Allen [2], p. 362.

[3] See Abbott [1], pp. 71–72 and Dorfman and Steiner [5], p. 833.

[4] Swan [7] has demonstrated that for some demand functions and cross-elasticities the monopolist would offer the same quality variety as would a group of competitors. But this need not be true for the general case See White [8], Chapter 11, for an argument that the American automobile industry has been slow in offering as much variety as would a more competitive industry.

Since an airline cannot charge for meals, its profit function is

$$\pi = P_T^* q_T - C_T q_T - C_M q_M q_T. \tag{3a}$$

The number of meals per passenger is the only uncontrolled variable in this situation. To find out how many meals a firm will offer per passenger, we take the first derivative with respect to q_M:

$$\frac{d\pi}{dq_M} = P_T^* \frac{\partial q_T}{\partial q_M} - C_T \frac{\partial q_T}{\partial q_M} - C_M q_M \frac{\partial q_T}{\partial q_M} - C_M q_T = 0. \tag{4}$$

This can be rearranged to

$$q_M = \frac{P_T^* - C_T}{C_M} - \frac{q_T}{\dfrac{\partial q_T}{\partial q_M}}. \tag{4a}$$

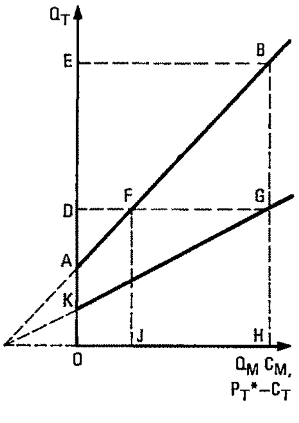
Equation (4a) gives the equilibrium quality level for a monopolist; but each firm among a group of competitors would see $\partial q_T / \partial q_M$ as being much larger than does the monopolist: the monopolist sees this term (correctly) as the overall market's response to an extra meal, while a competitor would imagine that by offering more meals he could steal transportation sales from the other competitors. The larger is the apparent $\partial q_T / \partial q_M$ to a competitor, the larger will be his quality offering. The members of a perfectly competitive industry would each see the demand elasticity to their own meals offerings as infinite, and $\partial q_T / \partial q_M = \infty$. (This is analogous to the infinite demand elasticity assumption in the usual price-quantity competitive model.) In this case, the last term of (4a) would disappear, and we are left with

$$q_M = \frac{P_T^* - C_T}{C_M}. \tag{5}$$

In words, each firm expands q_M in a competitive effort until zero profits are being made on each customer; i.e., until the cost of quality per customer $(q_M \cdot C_M)$ just equals the potential net revenue $(P_T^* - C_T)$ from basic transportation. Since only one price P_T^* can be charged in the industry, all competitive firms will have to offer the same quality level.

FIGURE I

BASIC MEALS MODEL



The model, then, indicates that a regulated monopolist will offer a lower level of quality per passenger than will a regulated competitive industry. The monopolist is aware that by offering an extra meal to the marginal customer he has to offer an extra meal to all inframarginal passengers (just as the price-quantity monopolist is aware of his marginal revenue curve). The competitor believes that the demand response to his extra quality offering will be so high that he ignores the inframarginal problem (as does the price-quantity competitor). Also, (4a) indicates that the more competitive is an oligopoly (as measured by the $\partial q_T / \partial q_M$ perceived by the oligopolists) the higher will be their quality offerings.

The model's results can easily be shown diagramatically. In Figure 1, curve AFB represents the response at a particular price P_T^* of overall transportation demand to quality costs per passenger incurred by firms. Since the unit costs of meals are constant, more expenditures per passenger by firms mean more meals per passenger and a higher market demand.[5] If P_T^* increases, curve AB will shift down, since at any quality level less transportation will be demanded at the higher price. Distance OH represents $P_T^* - C_T$, which is the potential net revenue from basic transportation.

---

[5] If the unit costs of meals change exogenously, curve AB will rotate on point A.

Competitive firms compete with each other by offering more meals per passenger. They do this until it is no longer profitable to spend more on meals; i.e., when they are each spending EB = OH on meals per passenger. This wipes out the OH potential net revenue per passenger that regulation temporarily created. Thus, OH/C_M meals per passenger are provided by each competitor, and the overall market demand for transportation is OE.

A nondiscriminating monopolist, concerned about the costs of extra meals to inframarginal customers, responds along curve KG, which represents the marginal cost of attracting transportation demand with an extra meal. He equates these marginal costs with the marginal net revenue per passenger, and thus he spends only DF = OJ on meals per passenger. Only OJ/C_M meals are provided per passenger, and only OD amount of transportation is demanded.[6] The monopolist's monopoly profits are represented by rectangle FGHJ.

A number of important conclusions emerge from this model:

(1) As we have seen, a competitive industry under price regulation will offer more quality per primary output unit than would a simple monopolist.[7]
(2) Consequently, a regulated competitive industry will sell more units of the primary output than will a regulated monopoly.
(3) The amount of quality per unit that will be provided by a competitive industry will be dependent on the price set by the regulator. The higher the price, the higher the quality per unit. This can be derived from (5), since

$$\frac{\partial q_M}{\partial P_T^*} = \frac{1}{C_M} > 0. \tag{6}$$

It is also clearly seen in Figure 1.

The monopolist's quality response to price changes is not so predictable. We have no *a priori* prediction for the way in which the response of market demand to quality level *changes* as the price of the primary output changes. If, as the price of transportation increases, the demand for transportation becomes less responsive to meals offerings, quality per unit might actually decrease rather than increase. In Figure 1, if curves AB and KG become flatter when HB shifts to the right, distance DF might conceivably become smaller. This can also be seen by taking total differentials in (4a) and then solving for dq_M/dP_T^*:

$$\frac{dq_M}{dP_T^*} = \frac{\dfrac{1}{C_M}\cdot\left(\dfrac{\partial q_T}{\partial q_M}\right)^2 - \dfrac{\partial q_T}{\partial q_M}\cdot\dfrac{\partial q_T}{\partial P_T^*} + q_T\cdot\dfrac{\partial^2 q_T}{\partial q_M \partial P_T^*}}{2\left(\dfrac{\partial q_T}{\partial q_M}\right)^2 - q_T\cdot\dfrac{\partial^2 q_T}{\partial q_M^2}}. \tag{7}$$

---

[6] If the monopolist were a discriminating monopolist, curve AB would be his marginal quality cost curve.

[7] This point has been noted empirically in a study of the airlines industry. See Caves [4], p. 424.

The denominator is positive, since the second-order condition for a maximum is that $q_T \cdot (\partial^2 q_T/\partial q_M^2) - 2(\partial q_T/\partial q_M)^2 < 0$. The sum of the first two terms of the numerator is positive, since $(\partial q_T/\partial q_M) > 0$ and $(\partial q_T/\partial P_T^*) < 0$; but the third term is unsigned. If this term is negative and large enough, the entire fraction could become negative.

(4) Surprisingly, a decrease in the regulated price may not generate an increase in the quantity demanded (and produced) of the regulated good. From (1a) we can derive

$$\frac{dQ_T}{dP_T^*} = \frac{\partial Q_T}{\partial P_T^*} + \frac{\partial Q_T}{\partial Q_M} \cdot \frac{\partial Q_M}{\partial P_T^*}. \tag{8}$$

For a competitive industry, the second term is positive and the first is negative. For a monopoly, the sign of the second term is indeterminate. In either case, there is no *a priori* way of determining the sign of dQ_T/dP_T^*. In terms of Figure 1, a decrease in P_T^* has two effects: it shifts curve AB up (at a lower price, more transportation will be demanded at any level of meals), and it shifts curve HB to the left (there is a smaller initial net revenue margin). The intersection of the two curves will move to the left, but we cannot tell whether it will move up or down. An extreme case may make this point clear: Suppose transportation demand is totally price-insensitive (AB does not shift when P_T^* changes) but is sensitive to meals. In this case, a decrease in P_T^* means fewer meals and smaller transportation demand.

(5) All purchasers will be no better off and most will be worse off under price regulation than under unregulated competition. This follows from the elimination of choice of how many meals a passenger can buy. Regardless of the price P_T^* set by the regulator, any passenger who previously wished to spend more or less on meals than $P_T^* - C_T$ must be worse off. Only the passenger who wanted to spend exactly $P_T^* - C_T$ on meals is equally well off under price regulation. A welfare comparison between monopoly with and without price regulation is more difficult. Passengers lose choice but gain a lower price for transportation.[8] No *a priori* judgment can be made.

(6) There is no way that the regulator can determine the "proper" or "optimum" regulated price P_T^* without knowing the detailed preference functions of the passengers and having an explicit interpersonal utility weighting function. Each P_T^* carries a different number of meals. Some passengers may prefer a high price and many meals, some may prefer a low price and few meals. Once the regulator chooses to regulate the price, there is no easy way out. The choice between regulated monopoly and regulated competition is easier. For any P_T^*, passengers are always better off under regulated competition: they get more meals. But is competition "wasteful" or "destructive"? Do the competitors provide "too many" meals, and would society's resources be better utilized by replacing them with a monopolist who would offer fewer meals? Again,

---

[8] Only if the regulator sets P_T^* above the preregulatory "average total price" $(P_T + P_M q_M)$ for the monopolist is everyone guaranteed of being worse off.

without a detailed knowledge of preference functions, no definite answer can be given.[9] In any event, the most efficient way to cut down on the offering of meals is not to monopolize the industry but to lower the P_T^* that the competitors can charge.

(7) From conclusion (6), it follows that a price-regulated competitive industry will almost always produce a smaller quantity of its primary output than it would without regulation. In the absence of some very strange preference functions, the decreased welfare from the forced price-and-meals package will lead to smaller demand for transportation. Again, for the case of monopoly, no *a priori* statement can be made.

(8) Finally, attempts by the regulator to give a competitive industry high profits by setting a high price will come to nought. All potential profits will be competed away in high per unit quality offerings.[10] Only the "normal" profits built into the cost functions will prevail.

The conclusions of this model have followed from the existence of an unregulated variable, meals, on which competitors could compete. Section 5 explores the possibility of allowing the price of meals to be the unregulated variable. But suppose the regulator also regulates the price of meals, by specifying a single P_M^* per meal that must be explicitly or implicitly charged by all firms. These results disappear. Market demand considerations based on P_T^* and P_M^* determine entirely how much transportation and how many meals are bought and produced. Competitors no longer have anything on which they can compete—unless a second unpriced quality variable is discovered. Market shares are determined by random choice or historical precedent, and the competitive and monopoly solutions are identical. Consumers have regained quality variety but at the cost of whatever excess profits are yielded by P_T^* and P_M^*.

4. The "advertising" model

■ A similar model can be constructed for "advertising," a quality variable that is not directly attached to each unit of transportation. The overall transportation demand function under price regulation can still be expressed as

$$Q_T = Q_T(P_T^*, Q_A) \tag{1b}$$

but the firm's profit function is now

$$\pi = P_T^* q_T - C_T q_T - C_A q_A. \tag{3b}$$

The derivative with respect to q_A is

$$\frac{d\pi}{dq_A} = P_T^* \frac{\partial q_T}{\partial q_A} - C_T \frac{\partial q_T}{\partial q_A} - C_A = 0 \tag{4b}$$

or

---

[9] This is equivalent to a Kaldor-Hicks compensation test: Can the monopolist bribe the passengers so they are equally well off with fewer meals, and still come out ahead? It depends on how strongly they like their meals. See also the discussion of this point in Caves [4], pp 423–24.

[10] I am indebted to George Eads for this point.

$$C_A = \frac{\partial q_T}{\partial q_A}(P_T^* - C_T). \qquad (4c)$$

In words, expand advertising until the cost of an extra unit of advertising equals the extra transportation demand it generates times the net revenue per transportation unit. This determines the amount of advertising for the monopolist.

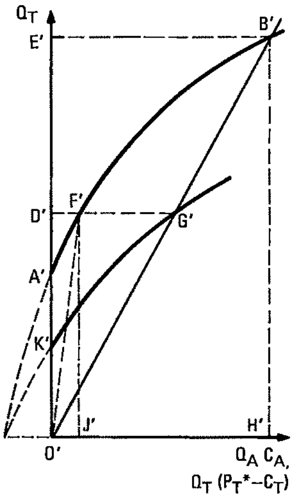
Each of a group of competitors would see advertising as a means of stealing sales and thus would see $\partial q_T/\partial q_A$ as being larger than does the monopolist. They will expand their advertising until zero total profits are being made, i.e., to the point where

$$C_A q_A = q_T(P_T^* - C_T). \qquad (5a)$$

But, unlike the case of meals, this zero profit point is not an automatic stopping point, except in the long run. In the short run, each marginal sale of transportation would still appear to be profitable since, unlike meals, advertising costs are not directly attached to transportation units. Each competitor would continue to be tempted to expand his advertising to try to steal transportation sales.[11] Thus, for short periods, q_A might be expanded beyond the zero profit point of (5a), and firms would run losses. Eventually, firms would have to retreat to q_A at the zero profit point. Thus (5a) would represent the long-run tendency for perfect competitors, while in the short-run there would be periodic expansions and contractions of q_A in the vicinity of this point.

FIGURE 2

ADVERTISING EXPENDITURE MODEL



We can easily demonstrate that the competitive industry's advertising (the sum of the individual competitors' advertising) is greater than the monopolist's advertising and the competitive advertising per unit of transportation is also greater than that of the monopolist. In Figure 2, curve A′F′B′ shows the response of transportation demand to advertising expenditure at a particular P_T. The curve must show diminishing returns, or else the amount of advertising would be expanded indefinitely. Line O′G′B′ represents the total net revenue from every level of transportation.

Perfect competitors will compete away their profits in advertising until they are spending E′B′ = O′H′. Transportation demand will be O′E′. The monopolist will follow his marginal cost curve, K′G′. This crosses curve O′B′ at G′ (where the tangent at F′ equals the slope of O′B′). He will spend only D′F′ = O′J′ on advertising, and only O′D′ transportation demand will be generated. He spends a smaller total amount on advertising and also a smaller amount per unit of transportation. This latter point is shown by comparing the slopes of line O′F′ and O′G′B′.

The eight conclusions derived from the meals model can equally well be derived from this advertising model.

5. The model with a separate price on quality

■ We can ask a further question about the basic meals model: What happens if the regulator specifies a fixed price for the basic output but leaves members of the industry free to charge any price they wish for the quality attributes?

---

[11] The overtime "investment" aspect of advertising is being ignored at this point. The temptation to expand advertising occurs solely because it appears to be a way of stealing current transportation sales.

We can use slightly modified forms of the original demand functions and profit formulation:

$$Q_T = Q_T(P_T^*, P_M), \tag{1c}$$

$$Q_M = Q_M(P_M, P_T^*) \tag{2c}$$

and

$$\pi = P_T^* q_T + P_M q_M q_T - C_T q_T - C_M q_M q_T. \tag{3c}$$

The variable open to competitive effort is P_M.[12] The derivative with respect to P_M is

$$\begin{aligned} \frac{d\pi}{dP} = \frac{\partial q_T}{\partial P_M}(P_T^* + P_M q_M - C_T - C_M q_M) \\ + \frac{\partial q_M}{\partial P_M}(q_T P_M - q_T C_M) + q_T q_M = 0. \end{aligned} \tag{4d}$$

Again, $\partial q_T/\partial P_M$ will appear to be very high or infinite to a group of perfect competitors. The other terms will shrink in importance,[13] and we are left, after re-arranging terms, with

$$q_M(C_M - P_M) = P_T^* - C_T. \tag{5b}$$

In words, each competitor continues to lower P_M in efforts to steal transportation sales from his competitors until it is no longer profitable to do so, i.e., the point at which the *loss* on meals $(C_M - P_M)$ per passenger offsets the net revenue from the transportation per passenger so that zero profits are being made on each combined package.[14]

But this zero profit situation is true only in an average sense; i.e., at the P_M implied by (5b), an *average* passenger will demand the corresponding q_M meals so that zero profits are being made. In fact, at price P_M a competitor would be making a combined package profit on those passengers who consumed fewer than the average number of meals and would be making a combined package loss on those who consumed greater than the average number of meals. Accordingly, provided that there was some way a firm could discriminate and segregate passengers by meal consumption (e.g., by having one-meal-only flights, two-meal-only flights, etc.), competition would force each firm to charge an even lower price than the P_M of equation (5b) to those passengers who wanted fewer than the average number of meals and a higher price (but still below C_M) to those who wanted more meals than average. Equilibrium would only be reached when different prices were charged to different passengers, such that

---

[12] Firms could try to offer more quality output for the same P_M, but this would be the equivalent of changing P_M, since we have defined P_M (and C_M) as a basic, unchanging unit.

[13] The term $(\partial q_M/\partial P_M)$ is likely to be much smaller than the apparent size of $(\partial q_T/\partial P_M)$. A drop in the price of meals can only induce more meals consumption per passenger from a firm's existing passengers, whereas the meals price drop will always be seen to be capable of stealing large numbers of passengers from other firms.

[14] The geometry of this proposition is basically similar to that of Section 3 and is not repeated here.

$$Q_{M,i}(C_M - P_{M,i}) = P_T^* - C_T, \tag{5c}$$

i.e., when the particular meals price charged to the *i*th passenger and the meals per passenger consumed at that price just yield a loss that offsets the net revenue from transportation. If a consumer wanted few enough meals, the competitive firms would pay him for the meal. (If this direct payment were not allowed, some indirect payment, like a positive value lottery ticket, might accompany a free meal.)

A numerical example may help clarify this point. Suppose $P_T^* = \$15$, $C_T = \$10$, and $C_M = \$2$. Then competition would force each firm to offer the meals price schedule of Table 1.

TABLE 1

MEALS PRICE SCHEDULE WITH COMPETITION

TYPE OF FLIGHT	PRICE PER MEAL	COST OF MEAL (S)	TOTAL MEAL DEFICIT
0–MEAL FLIGHT	–$5.00	$0	$5
1–MEAL FLIGHT	– 3.00	2	5
2–MEAL FLIGHT	– 0.50	4	5
3–MEAL FLIGHT	0.33	6	5
4–MEAL FLIGHT	0.75	8	5
5–MEAL FLIGHT	1.00	10	5
ETC.			

Thus regardless of the number of meals desired, a consumer will always pay $5 less for his total number of meals than he would in the absence of regulation (since in the absence of regulation, $P_M = C_M$). This $5 will always offset the regulated price increase. Consumers will be induced to buy the same number of flights and meals as they would in the absence of regulation. As a final confirmation of this point, close scrutiny of Table 1 will reveal that the marginal cost to a consumer of buying an extra meal is always $2, the same as it would be in the absence of regulation.

Accordingly, every consumer will be as well off under this form of regulation, as long as competitive firms can discriminate, as he would be in the absence of regulation. *And it will not matter what regulated price* P_T^* *is set by the regulator.* Any excess of P_T^* over C_T will always be given back to the consumer in a variable meals package which the consumer can choose. Thus the outcome of this form of regulation is the same as that of unregulated competition. The nominal P_T and P_M may be different under this form of regulation than under unregulated competition,[15] but the basic outcomes of Q_T, Q_M and consumer welfare will be identical in both cases. *The regulator can have absolutely no effect on the outcome.*

This is a surprising and not entirely obvious result. Under unregulated competition, firms were under no pressure to discriminate and would sell meal tickets at the door. Under price regulation and no charge for meals (Section 3), firms could not discriminate and offered the same number of (free) meals to everyone. Here, under price regulation and a charge allowed for meals, competition forces each firm to try to discriminate among its customers; but it is this discrimination that ensures the same outcome as unregulated competition.

---

[15] And, as the nominal P_T^* goes up, the nominal $P_{M,i}$ falls. From (5c) we can derive $(\partial P_{M,i}/\partial P_T^*) = -(1/q_{M,i}) < 0$

What happens if firms are not able to discriminate or are required by the regulator to charge one meals price to all customers? If firms can set maximum limits on meals consumption and if different firms can set different meals prices (though once it chooses a price, each firm can charge only that price to all of its customers), the previous conclusions still hold. We would see the firms in the industry turn to specialization. One-meal airlines, 3-meal airlines, 6-meal airlines, etc., would evolve, each being forced by competition to charge the appropriate zero total profit price for meals. Consumers would face the same alternatives as if each firm could discriminate.

If, however, one industry-wide price for meals is required by the regulator (though the forces of competition are allowed to determine that one price), an outcome identical to that of no-charge-for-meals regulation (Section 3) will evolve. The reasoning here is as follows. Competition forces the single price to yield an average loss on meals which just offsets the net revenue from basic transportation, so that an average of zero profit is made per passenger, as shown in (5b). But firms will notice that they are losing money on passengers who consume more than the average number of meals. If a firm cannot raise the price of meals to those customers, it can limit the number of meals they can consume. Since these customers are unprofitable, there would be no competitive barrier to limiting their meals consumption. But any effective limit reduces these losses and generates overall profits. The "average" extra customer again appears profitable to the firm. Competition again forces down the price of meals until zero profits are made. Again, the above-average meals consumers are now causing losses, and so on. This process will continue as long as a positive price is charged for meals. The logical stopping place is a price of zero, at which point competition forces firms to set a maximum limit on meals consumption equal to the net revenue margin on transportation divided by the cost of a meal. Since the price is zero to the customer, it is likely that all customers will take advantage of the maximum, so that the average and the maximum coincide. But this condition leads us to result (5) and all of the conclusions of Section 3 would follow.

A monopolist under this charge-for-meals regulation will follow (4d) fully. For the same P_T^*, he will charge a higher price for meals than the average price charged by competitors. If he can discriminate, each of his meals prices will be higher than the comparable one set by competitors. Again, consumers will clearly be better off under competition than under monopoly. Finally, one could easily demonstrate the ambiguities of the monopolist's transportation response and meals response to changes in P_T^*, since, again, we have no *a priori* predictions for the signs of $\partial^2 q_T/\partial P_M \partial P^*{}_T$ and $\partial^2 q_M/\partial P_M \partial P^*{}_T$.

References

1. ABBOTT, L. *Quality and Competition.* New York: Columbia Univ. Press, 1955.
2. ALLEN. R. G. D. *Mathematical Analysis for Economists.* London: Macmillan & Co., 1962.
3. ARCHIBALD, G. C. "Profit-Maximizing and Non-Price Competition." *Economica,* Vol. 31, No. 121 (February 1964), pp. 13–22.
4. CAVES, R. E *Air Transport and Its Regulators.* Cambridge, Mass.: Harvard Univ. Press, 1962.
5. DORFMAN, R. AND STEINER, P. O. "Optimal Advertising and Optimal Quality." *The American Economic Review,* Vol. 44, No. 5 (December 1954), pp. 826–36
6. STIGLER, G. J. "Price and Non-Price Competition." *Journal of Political Economy,* Vol. 76, No. 1 (January/February 1968), pp. 149–54.
7. SWAN, P. L. "Market Structure and Technological Progress: The Influence of Monopoly on Product Innovation." *Quarterly Journal of Economics,* Vol. 84, No. 337 (November 1970), pp. 627–38.
8. WHITE, L. J. *The Automobile Industry Since 1945.* Cambridge, Mass.: Harvard Univ. Press, 1971.

Models of the demand for air transportation

Philip K. Verleger, Jr.
Senior Research Economist
Data Resources, Incorporated

Due to the elaborate regulatory procedures established by the U. S. Civil Aeronautics Board, the demand for transportation by air has been the subject of many studies. The purpose of this paper is to compare three different empirical approaches for suitability to the analysis of demand. Two aggregate and one disaggregate approaches are explored. In addition, a more general method of measuring mass in gravity models is introduced. The results from the disaggregated model indicate the inadvisability of using aggregate or cross-sectional models for the analysis of the effects of regulatory policy.

1. Introduction

■ Empirical analyses of demand for transportation and communication have appeared from time to time in specialized journals. Frequently, the models examined have been very abstract and often unsupportable by the fragmentary data employed to test them. This paper attempts to deal with the problem in a different way. The demand for a single transportation service, air transport, is analyzed within three different relatively simple theoretical frameworks, cross-sectional, aggregate, and point-to-point. The purpose of examining demand in this manner is to determine whether any or all of the methods yield consistent estimates of demand. As a by-product, a slightly different specification of the mass variable in the gravity model which has been widely used in this type of demand analysis is introduced.

This paper is organized into five sections. Section 2 is devoted to a discussion and a test of the assumption behind a cross-sectional model of demand for air transportation. Aggregate models are discussed, analyzed, and tested in Section 3. Section 4 is devoted to the analytical specification of the point-to-point model, with particular emphasis on the construction of the mass variable. Finally, Section 5 presents an analysis of the results from the estimates of the point-to-point model over 115 city pairs.

2. Gravity and cross-sectional models of transportation

■ The gravity model is a variant on the gravitational law of physics discovered by Newton. In the abstract, it assumes that travel

---

Philip K. Verleger, Jr. received the A.B. degree from Cornell University (1966) and the Ph.D. degree in economics from M.I.T. (1971). He is a senior research economist at Data Resources and specializes in microeconomic applications of econometrics.

This paper was developed primarily out of the author's thesis at M.I.T. He is indebted to F. M. Fisher, P. W. MacAvoy, and R. M Solow for assistance with the many drafts of his work. He owes special thanks to Professor MacAvoy for encouraging him to write this paper, and to an anonymous referee.

or communication between two cities will decrease as the distance between them increases. Since cities are stationary, this principle has been applied to cross-sectional models of travel. Travel between a set of cities has been supposed to be directly proportional to the population of the two cities, and inversely proportional to distance squared. That is, travel between cities i and j is given by

$$T_{ij} = a\frac{M_i M_j}{d_{ij}^2}, \tag{1}$$

where M_i and M_j are the populations of cities i and j, and d_{ij} is distance. The simple model may, however, be varied in a number of ways. One variation, given by

$$T_{ij} = a\frac{M_i^{\alpha_i} M_j^{\alpha_j}}{d_{ij}^2}, \tag{2}$$

drops the assumption that populations of each city have equal travel inducing effects. An even more generalized gravity model has been proposed by Quandt:[1]

$$T_{ij} = M_i^{\alpha_1} M_j^{\alpha_2} Y_i^{\alpha_3} Y_j^{\alpha_4} V_i^{\alpha_5} V_j^{\alpha_6} H_i^{\alpha_7} C_i^{\alpha_8} S_{ij}^{\alpha_9}, \tag{3}$$

where the variables are as before: Y represents income; V, some characterization of the city type such as by percent of labor employed in manufacturing and mining; H and C, cost and travel time variables; and S, a variable for supply convenience of travel.

The gravity model specification is most frequently tested against cross-sectional data. It has often been used to determine, or attempt to determine, personal preference for a mode of travel when faced with two or more alternatives.[2] While these models are very complex and frequently require the use of very complicated estimation techniques, they always require the assumption that travel across a diverse set of city pairs can be characterized by the same set variables. The Mathematica studies[3] have normally used a set of similar cities, such as those in the Northeast Corridor from Boston to Washington, for which that assumption may have some chance of being fulfilled. However, there have been many other studies for which such assumptions probably will not hold.

One such study, constructed by the Civil Aeronautics Board staff in connection with a recent domestic passenger investigation, gives an opportunity to examine the assumptions.[4] The CAB study presents a reestimation of a model developed by Brown and Watkins in 1966.[5] The model is specified as a form of the gravity model

$$T_{ij} = aP_{ij}^{\gamma} M_{ij}^{\beta} C_{ij}^{\mu} D_{ij}^{\delta} (ET)_{ij}^{\zeta}, \tag{4}$$

---

[1] See [9].

[2] Mathematica has published four volumes of such studies, *Studies in Travel Demand* [8], done in connection with the Department of Transportation's Northeast Corridor Project. These works are generally based on the work of Quandt and Baumol [10].

[3] *Ibid.*

[4] See [15].

[5] See [2]

where

P = price,
M = the product of aggregate income in i times that in j,
C_{ij} = phone calls between i and j,
D_{ij} = distance between i and j, and
ET_{ij} = elapsed flying time between i and j.

The CAB staff estimated the model over 441 city pairs ranging in distance from 106 miles (Hartford to New York) to 2691 miles (Boston to San Francisco), including both very large markets (New York to Washington) and very small markets (New York to Sacramento). While the model clearly was naive in its assumption that phone calls were a measure of community interest and in assuming away the presence of competition by other modes, it offered a chance to test the stability of coefficients estimated by simple gravity models with respect to distance and market size.[6] The model was reestimated by the present author for all city pairs and then for city pairs in successive 500 mile intervals (see Table 1). In spite of the presence of the phone call variable, ordinary least squares was used throughout.[7]

TABLE 1

CROSS–SECTIONAL RESULTS LISTED BY DISTANCE*

DISTANCE (MILES)	CONST	PRICE	INCOME	CALLS	DIST.	TIME	R^2	OBS	SSR**
ALL CITY PAIRS	8.92 (4.21)	−0.92 (3.42)	0.14 (4.62)	0.52 (15.44)	−0.32 (3.96)	−0.40 (3.45)	0.72	441	101.15
0 TO 500	5.64 (1.84)	−1.03 (2.84)	0.12 (2.92)	0.54 (8.88)	0.03 (0.15)	0.21 (0.79)	0.65	100	18.01
500 TO 1000	9.32 (1.90)	−0.91 (1.55)	0.18 (3.59)	0.47 (8.60)	−0.20 (0.81)	−0.79 (3.71)	0.70	129	19.74
1000 TO 1500	11.17 (1.67)	−2.08 (2.59)	−0.32 (2.53)	0.99 (7.98)	0.35 (0.69)	−0.80 (0.44)	0.67	91	22.29
1500 TO 2000	7.33 (0.73)	1.03 (0.99)	0.35 (3.22)	0.18 (1.76)	−1.24 (1.57)	−1.60 (3.26)	0.64	61	13.66
OVER 2000	23.09 (2.80)	−1.62 (2.20)	−0.02 (0.14)	0.70 (6.21)	1.31 (1.24)	−3.77 (6.01)	0.83	60	10.16

*t STATISTICS IN PARENTHESES.
**SUM OF SQUARED RESIDUALS.

---

[6] This model was originally presented by Brown and Watkins [2]. This particular study was presented in the direct testimony of Wayne Watkins during the discount fare phase of the *Domestic Passenger Fare Investigation* [15].

[7] This is clearly an example of a case where simultaneous methods should be used, because phone calls and airline traffic are probably subject to the same stochastic errors. For the entire sample, ordinary least squares and two-stage least squares produced the following results (t statistics are in parentheses):

	a	P	M	Phone	Dist	Time	R^2
OLS	8.92 (4.21)	−0.92 (−3.42)	0.14 (4.62)	0.52 (15.44)	−0.32 (3.96)	0.40 (3.45)	0.71
TSLS	−1.32 (0.10)	−1.81 (0.55)	−0.36 (0.55)	1.30 (1.23)	2.22 (0.85)	0.04 (0.55),	0.37

with the price and phone calls treated as endogenous variables in the second equation. The model clearly collapsed in the second case, as it probably should. Simultaneous estimates of equations in smaller distance classes produced even more disastrous results.

The results of Table 1 indicate that the coefficients are subject to large variations. The price coefficient varies from 1.03 to −2.08 whereas the income coefficients between distance groups are different at the 1-percent level.[8] As one would suspect, demand functions for travel by a single mode vary with distance, thus preventing the direct application of a gravity model to a single sample.

As a final experiment, the CAB data were segregated by market size and the simple model was tested against only those city pairs with relatively heavy traffic in the sample period.[9] Again the coefficients changed dramatically. See Table 2.

These results tend to put the cross-sectional gravity model in a bad light, but this is an incorrect interpretation. Gravity models have produced very satisfactory results for homogenous sets of city pairs (such as the Northeast Corridor). However, taken together, the U. S. air travel market is composed of a very diverse set of heterogenous markets which cannot be easily lumped together. One can conclude from this analysis that it is impossible to use cross-sectional techniques to analyze the entire air transportation market.

TABLE 2

CROSS–SECTIONAL RESULTS FOR CITY PAIRS WITH HEAVY TRAFFIC *

	FULL SAMPLE	MAJOR CITY PAIRS
CONSTANT	8 92 (4.21)	12.93 (5.28)
PRICE	−0.92 (3.42)	−1.74 (5.61)
INCOME	0.14 (4.62)	0.04 (1.05)
PHONE CALLS	0 52 (15.44)	0.52 (11.40)
DISTANCE	−0 32 (3.96)	−0.15 (−1 52)
TIME	−0.40 (3.45)	0.078 (0.35)
R^2	0.72	0.65

*t STATISTICS IN PARENTHESES.

3. Aggregate travel models

■ In the literature on the demand for travel, particularly by air, the aggregate model appears most frequently. This model treats all air travel, usually measured in revenue passenger miles, as a homogenous commodity[10] and bears a strong kinship with many other aggregate demand functions such as for copper. However, while models for commodities have a strong connection to some underlying microeconomic relationship, the aggregate travel demand model has almost no connection to the underlying microtravel demand functions. This model has been useful for accounting purposes and revenue forecasting. It is not, however, comparable to aggregate measures of a commodity (such as tons of No. 3 winter wheat). In fact, an aggregate measure of travel demand is more like a measure of gross output of a group of commodities,

---

[8] The Chow F test given by

$$F = \frac{S^2 - \sum_{i=1}^{5} S_i^2}{N - \sum_{i=1}^{5} N_i} \quad \frac{\sum_{i=1}^{5} S_i^2}{\sum_{i=1}^{5} N_i}$$

where

S = the sum of squared residuals from the regression on the entire sample,

S_i = the sum of squared residuals from the i^{th} subregression,

N = the degrees of freedom in the main regression, and

N_i = the degrees of freedom in the i^{th} regression,

has the value 3.76 while $F(24,411,0.01) = 1.84$.

[9] The cutoff was set at 100 passengers. This reduced the sample from 441 observations to 171. By comparison, the city pair with the heaviest traffic, Miami to New York, had 2,000 passengers.

[10] A revenue passenger mile is defined as one paying passenger moved one mile. Thus, every person flying from Boston to New York generates 188 revenue passenger miles (RPM's) while every passenger from New York to Los Angeles generates 2,400 RPM's.

such as the gross output of all agricultural production measured in tons. In this section, we will discuss the problems involved in interpreting results from aggregate functions and models of demand, and also indicate an intrinsic problem occurring in these models due to the use of an average revenue variable called "yield per passenger mile" to represent price.

In most aggregate works[11] the aggregate measure of demand is postulated to be a function of some measure of price, a measure of average national income, and perhaps some measure of an alternative method of travel:

$$\mathrm{RPM}_t = aY_t^{\beta}(AR_t)^{\eta}\epsilon, \tag{5}$$

where Y_t represents per capita disposable income and AR_t, yield per passenger mile. The price variable chosen is usually a measure of average revenue:

$$AR_t = \frac{\sum_{i=1}^{n} T_t^i F_t^i}{\mathrm{RPM}_t}, \tag{6}$$

where T_t^i represents traffic on route i and F_t^i represents the average fare on route i.[12] The use of average revenue as a price surrogate prevents the interpretation of the aggregate function as a demand function, except in the special case where all prices are proportional over time.[13]

While the bias this procedure imposes on the aggregate model can be shown mathematically,[14] the problem can be more easily explained by example. Suppose there are two routes and route A has a lower price per mile but a higher income elasticity than B. If, over time, prices never change, but income increases, the average revenue per passenger mile will decline. If this variable were used in an aggregate demand function, it is quite possible that a statistically significant price coefficient might be estimated *in spite of the fact that no price experiment was ever conducted.*

To demonstrate the applicability of this example, in Figure 1 the yield per domestic passenger mile is compared with a very rough price index constructed by the author. The index was prepared using data for only three routes, New York to Los Angeles, New York to Chicago, and Chicago to Los Angeles, using fixed weights representing the proportion of first-class and coach traffic on each route in 1967. Its major weakness is that discount fares are excluded (this primarily affects the period after 1966). The fundamental variation between the price index and yield per passenger mile can be traced to three items: shifting the mix from first class to coach; the change in average stage length; and the introduction of discount fares. The impact of the use of a price index rather than yield per passenger mile can be measured by comparing the

---

[11] See Bartlett [1], Brown and Watkins [2], Institute for Defense Analysis [6], and Joun [7].

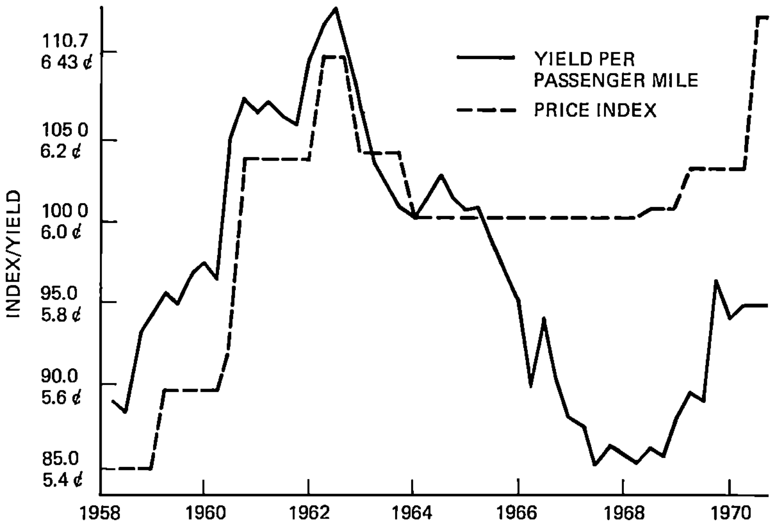
[12] Average fare is again an average revenue variable. It is the average fare paid by all passengers on a given route.

[13] Clemins [4] has developed a simple explanation of this bias introduced when average revenue is used in place of a price index

[14] See Verleger and Eads [18].

FIGURE 1

YIELD PER PASSENGER MILE COMPARED WITH A SIMPLE AIRLINE PRICE INDEX



results of regression using each price variable. These results are presented in Table 3 for the interval 1958–1965. This interval was selected to minimize the impact of discount fares after 1966.

The results indicate that equations using average revenue variables have higher price elasticities than those estimated using price indexes. This seems to be the natural result on *a priori* grounds since the experience from 1958 to 1970 was one of increasing income, increasing trip lengths, and increasing availability of coach seats. The increase in trip length would reduce average revenue per passenger mile due to the decrease in price per mile with increasing distance. The increase in coach travel would also tend to lead to a reduction in average revenue. Some of this error may be correctable by the use of full information methods. If the average revenue variable is treated as endogenous and two-stage least squares used, the differences between results obtained using the price index and average revenue variable are greatly reduced, as is apparent from Table 4. There the price index results are identical to those given in Table 3 because the index is treated as an exogenous variable.[15]

Much of the difference between average revenue and price index approaches can be traced to simultaneous equation bias.[16] This suggests that airline prices may in fact have moved in a sufficiently uniform manner that average revenue can be used as a price index if corrected for shifts in the mix between first-class and coach travel, and if the average trip length held constant. However, aggregate equations, even correctly constructed, can continue to predict accurately only if demands between individual city pairs remain relatively identical. In the next section we shall investigate the similarities of demand across city pairs.

TABLE 3

ORDINARY LEAST-SQUARES REGRESSION RESULTS: AVERAGE REVENUE VS. PRICE INDEX, 1958–1965*

	PRICE VARIABLE	
	AVERAGE REVENUE	PRICE INDEX
METHOD OF ESTIMATION	OLS	OLS
CONSTANT	−0.848 (2.069)	−0.513 (10.875)
PRICE @ COEFFICIENT	−0.122 (0.366)	−0.053 (0.671)
INCOME** COEFFICIENT	2.370 (38.470)	2.385 (33.471)
R^2	0.98	0.98
DURBIN-WATSON	1.14	1.14

*t STATISTICS IN PARENTHESES.

**INCOME IS PERSONAL DISPOSABLE INCOME PER CAPITA LESS TRANSFERS

@PRICES ARE IN CURRENT DOLLARS CONSTANT DOLLAR ESTIMATES PRODUCED SIMILAR RESULTS.

---

[15] The price index is treated as exogenous due to the nature of the CAB's administration of price, which involves very long lags. Thus, current revenues and cost may affect next year's prices, but not today's. Income is treated as an exogenous variable throughout

[16] For some reason no author of an aggregate model has ever attempted to test for simultaneous equation bias although our results clearly indicate its presence.

4. City pair models of the demand for air transportation

■ Cross-sectional and aggregate models can provide estimates of average values of parameters. Theil has shown[17] that these estimates may not be unbiased, and the analysis above indicates that they may not even have small variances. For regulatory policy, it is important to know whether the true parameters bear any relationship to the aggregate averages.

The CAB has collected point-to-point traffic data on a sample basis since 1950. These data make it feasible to test the stability of coefficients across city pairs. The remainder of this paper is dedicated to the test of such a hypothesis. The model developed treats total air travel between each pair of cities without regard to the composition of travel (as between business and pleasure travel), general economic conditions, or travel on other city pairs. In addition to allowing a test of the equality of coefficients across routes, the model provides as a by-product a test of the specification of mass variables in gravity models. In the following subsection a general model of point-to-point demand is developed.

TABLE 4

TWO-STAGE LEAST SQUARES REGRESSION RESULTS AVERAGE REVENUE VS PRICE INDEX, 1958–1965*

	PRICE VARIABLE	
	AVERAGE REVENUE	PRICE INDEX
METHOD OF ESTIMATION	TSLS	TSLS
CONSTANT	–0.333 (0.678)	–0.513 (10.875)
PRICE @ COEFFICIENT	0.056 (0.330)	–0.058 (0.671)
INCOME** COEFFICIENT	2 351 (3.6442)	2.385 (33.471)
R^2	0.98	0 98
DURBIN–WATSON	1.10	1.14

*t STATISTICS IN PARENTHESES.

**INCOME IS PERSONAL DISPOSABLE INCOME PER CAPITA LESS TRANSFERS.

@PRICES ARE IN CURRENT DOLLARS. CONSTANT DOLLAR ESTIMATES PRODUCED SIMILAR RESULTS.

☐ **Models of demand.** The choice of a measure of demand is difficult. Although the analysis which follows uses point-to-point traffic to represent demand, it does not follow that such data represent the best measure of true demand. Each observation of total traffic between two points represents the sum of individuals traveling for many different purposes. It may certainly be possible that a better estimate of demand can be obtained by considering only those persons traveling for a specific purpose than by considering all persons traveling from point A to point B. It may even be the case that aggregate measures of demand for travel by particular purposes or by certain corporations may provide better estimates of demand than disaggregate point-to-point models. However, data for the estimation of point-to-point travel models are readily available, while data for the estimation of travel demand by corporations are not. Thus the analysis which follows will attempt to deal with total air travel between isolated points. The aim of this work is to develop a model which should indicate whether differences in demands exist across city pairs, whether there is any regularity to these differences, and thus whether cross-sectional or aggregate models are theoretically feasible.

Specification of a point-to-point model of travel would seem simple in abstract terms in light of the earlier discussion. Travel should be some function of price, income, and perhaps certain other variables relating to changes in the availability of other modes of travel. In this work a modified form of the gravity model discussed earlier is selected. Although the gravity model may not always be suitable for the cross-sectional analysis of the demand for travel because of the intrinsic differences in the attractiveness of different cities, it offers a means of analyzing travel over time between two points. In the simplest form the model is specified as

$$T_{ij}(t) = aP_{ij}(t)^{\alpha}M_i(t)^{\beta_1}M_j(t)^{\beta_2}\epsilon, \qquad (7)$$

where $T_{ij}(t)$ is travel at time t between i to j; $P_{ij}(t)$ is the price prevailing at t; $M_i(t)$ is the mass at i at t; and $M_j(t)$ is the similar measure

---

[17] In [12].

of mass at j at t. The mass variable can be measured either by population or some population variable augmented by the propensity to travel of each segment of the populace. Thus mass might be measured by total population, per capita income, or some combination which augments population by the propensity to travel ($g_i^K(t)$) of individuals. Then mass would be given by a function

$$M_i(t) = \sum_{K=1}^{N} X_i^K(t) g_i^K(t), \tag{8}$$

where X_i^K represents the K^{th} individual (or group of individuals) residing in i and g_i^K represents his (or their) propensity to travel. In such a model $M_i(t)$ equals the population of i if the $g_i^K(t)$ are assumed to be uniformly equal to unity.

The measure of mass given in (8) seems to be more appropriate to the gravity model of travel demand, since travel increases with income. A change of a given magnitude in this variable should always cause the same change in travel, just as a given change in the magnitude of the mass of one physical body will always cause the same change in the gravitational attraction of that body to another. The simple population measure of mass does not meet this criterion unless the consumption of travel is invariant to changes in the level of income.

The method employed in this work is to assume that the function $g_i^K(t)$ is distributed exponentially:

$$g_i^K(t) = e^{\beta \bar{Y}_i^K(t)}, \tag{9a}$$

where $\bar{Y}_i^K$ is individual K's income (or average income for the group). Then M_i is given as

$$M_i = \sum_{K=1}^{N} X_i^K(t) e^{\beta \bar{Y}_i^K(t)}. \tag{9b}$$

Using this specification, it is possible to test the usual assumptions which are contained in all the works discussed above: that travel either (1) is distributed independently of income, or (2) increases proportionately with income. The test of assumption (1) is given by examining the estimate of $\beta : \hat{\beta}$. If $\hat{\beta}$ is not different from zero, the assumption is accepted; otherwise it is rejected. The test of assumption (2) (which is always made when an average income per capita term or aggregate income is included in a log linear equation), also depends on the relationship between β and the income distribution, since travel in such models as (3) and (4) is invariant to shifts in income distribution that leave aggregate income and income per capita unchanged. Travel in models with mass variables specified as in (9b), however, is clearly not invariant to income distribution shifts when $\beta \neq 0$.

☐ **A testable hypothesis.** Two models were tested using this hypothesis for 115 city pairs. The basic model tested a very simple nonlinear gravity model. Travel was made a function of price, mass variables, and seasonal dummies. Two possible structures of this model can be estimated. The first is given by

$$T_{ij} = \alpha P^{\beta} \frac{(M_i M_j)}{d_{ij}^2} \epsilon. \tag{10}$$

The second is given by

$$T_{ij} = \frac{\alpha P^{\beta}(M_i M_j)^{\gamma}}{d_{ij}^2}\epsilon. \tag{11}$$

The results reported in this paper are for just the second assumption, although (11) would seem to be more ad hoc than (10). Equation (11) allows a better test of the assumption that $\beta = 0$ than (10) because the restriction $\gamma = 1$ causes an apparent upward bias on β. The relationship between γ and β is discussed below.

Estimation of the demand function for a large number of city pairs was not an easy task. One major problem occurred frequently with those routes where service had not been adequate for the entire interval. On such routes, results were very erratic. Table 5 amplifies this problem for four routes. On two routes, New York to Los Angeles and New York to Miami, coach service was apparently available throughout the period. On two others, New York to Louisville and Cincinnati to Philadelphia, coach service was

TABLE 5

TRAFFIC FOR FOUR SELECTED ROUTES, 1960–1967

YEAR	QTR.	LAX NYC			MIA NYC			NYC LOU			CIN PHL		
		FIRST	COACH	TOTAL	FIRST	COACH	TOTAL	FIRST	COACH	TOTAL	FIRST	COACH	TOTAL
1960	1	4104	6290	10394	6613	24353	30966	1080	235	1315	479	9	488
	2	4108	8780	12888	3517	18917	22434	1447	293	1740	535	19	554
	3	3835	10490	14325	1970	17119	19089	1360	316	1676	471	5	476
	4	3275	6487	9762	3057	17579	20636	1200	257	1457	489	3	492
1961	1	3275	6034	9309	4004	23296	27300	1060	200	1260	468	10	478
	2	3194	8479	11673	2506	20025	22531	1295	337	1632	523	28	551
	3	3024	10370	13394	1444	17421	18865	1005	523	1528	598	28	626
	4	2818	8016	10834	2634	20098	22732	1124	560	1684	442	21	463
1962	1	2481	6976	9457	3723	27184	30907	736	582	1318	498	18	516
	2	2427	9567	11994	2342	15448	17790	912	975	1887	505	28	533
	3	2141	9976	12117	1262	16127	17389	569	965	1534	429	35	464
	4	2356	9289	11645	2085	17338	19423	700	1329	2029	500	56	556
1963	1	2276	8200	10476	3575	26211	29786	549	978	1527	435	67	502
	2	4086	9915	14001	3192	19440	22632	828	1317	2145	521	108	629
	3	4342	10564	14906	2324	18072	20396	782	1221	2003	338	98	436
	4	3498	8152	11650	4071	17893	21964	712	1472	2184	386	143	529
1964	1	3001	7556	10557	7936	25305	33241	506	1263	1769	274	142	416
	2	4236	13010	17246	3903	20303	24206	694	1793	2487	373	205	578
	3	4146	15064	19210	1816	23674	25490	645	1878	2523	397	247	644
	4	4320	10033	14353	3383	21569	24952	553	1754	2307	416	214	630
1965	1	4411	8999	13410	5352	32288	37640	508	1266	1774	439	216	655
	2	4635	14191	18826	3379	26506	29885	677	1948	2625	428	271	699
	3	4378	17022	21400	2252	25934	28186	597	1821	2418	362	414	776
	4	4611	11182	15793	4517	28127	32644	653	1744	2397	451	441	492
1966	1	4611	10460	15071	5511	33780	39291	488	1653	2141	436	424	860
	2	6092	15440	21532	3654	27413	31067	721	2347	3068	560	448	1008
	3	5068	18883	23951	2340	17797	20137	653	1988	2641	288	349	637
	4	5130	13766	18896	4746	30893	35639	671	2206	2877	288	796	1084
1967	1	5074	12151	17225	6402	43212	49614	537	2081	2618	200	826	1026
	2	5116	18393	23509	3725	31498	35223	753	2716	3469	179	978	1157
	3	4817	24747	29564	2754	33285	36039	528	2662	3190	128	1017	1145
	4	5456	16860	22316	5005	36984	41989	627	2342	2969	173	1119	1292

SOURCE: ORIGIN–DESTINATION SURVEY OF DOMESTIC AIRLINE PASSENGER TRAFFIC [14].

clearly rationed.[18]

5. Results from the estimation of the point-to-point model

■ Several specifications of the demand model were tested. Basically, a simple log linear demand model

$$T_{ij}(t) = \alpha_0 P_{ij}(t)^{\alpha_1} \tilde{Y}_{ij}(t)^{\alpha_2} \epsilon, \tag{12}$$

where $\tilde{Y}$ measures average per capita income, was compared to various specifications of the gravity model. The log linear demand model performed better in initial estimates than the gravity model. However, a modified form of the gravity model, as given by

$$T_{ij}(t) = \alpha_0 P_{ij}(t)^{\alpha_1} \left\{ \left(\sum_{K=1}^{N} X_i^K(t) e^{\beta_i \bar{Y}_i^K(t)} \right) \left(\sum_{K=1}^{N} X_j^K(t) e^{\beta_j \bar{Y}_j^K(t)} \right) \right\}^{\gamma} \epsilon \tag{13}$$

performed better on the average than the log linear model. The results indicate that (13) can be estimated as well with $\beta_i = \beta_j$ as with β_i and β_j allowed to vary separately. In fact, it is often impossible to estimate β_i and β_j separately, since the Hessian of the likelihood function becomes singular.

The gravity model as modified above was estimated for 115 city pairs. While the results from these estimations, given in Table 6, are not so satisfactory as had been hoped, they demonstrate conclusively that the modified special construction of the mass variable is significant. In almost every case equal weights for all individuals which would be obtained if the coefficient β equaled zero were rejected in favor of increasing weights, $\beta > 0$. The estimates of the price coefficient were not so regular, although those routes with better service in 1960 yielded better estimates of the price coefficient than those with poor service. Most of the point estimates of the price coefficients indicate that demand is somewhat price inelastic, but the standard error of the coefficients is frequently sufficiently large that this conclusion is very weak.

The model estimated for all routes was the one given by

$$T_{ij} = \alpha P_{ij}(t)^{\alpha_1} (M_i(t) M_j(t))^{\gamma}$$

$$\exp(\alpha_1 \text{SEA}\,1 + \alpha_2 \text{SEA}\,2 + \alpha_3 \text{SEA}\,3)\epsilon, \tag{14}$$

where M_i and M_j are given by

$$M_i = \sum_{K=1}^{N} X_i^K(t) e^{\beta_i \bar{Y}_i^K(t)}$$

and SEA represents a 0,1 seasonal dummy.

The price variable (P) represents an index formed using fixed weights which reflected passenger usage of the major airline fares: first class, coach, family plan (coach and first class), and discount

---

[18] The shift in the amount of coach travel results from the rationing of coach seats caused by the airlines' configuration of their airplanes. American Airlines' service to upstate New York offers a classic example of this problem. When service was performed with Lockheed Electra turboprop planes, 17 coach seats and 64 first-class seats were available on every flight. When service was shifted to Boeing 727 jets, 64 or 96 coach seats were available. Thus in the fourth quarter of 1964, when no jet service was available, 80 percent of the passengers from Buffalo to New York flew first class. By the final quarter of 1967 when most of the service was turbojet equipment, only 30 percent of the traffic was in first class. The New York-Rochester route showed a similar change from 89 percent first class in 1964 to 43 percent in 1967.

TABLE 6

ESTIMATES OF THE DEMAND FOR AIR TRANSPORTATION: ESTIMATION PERIOD, 1960–1967*

CITY PAIR	CONST.	PRICE	SEA 1	SEA 2	SEA 3	γ	β	INCOME ELASTICITY	R^2	DURBIN–WATSON	DISTANCE	PASSENGER RANK (1959)**	PASSENGERS 1959**
NYC PHL	–14.23 (12.39)	–0.74 (0.60)	–0.20 (0.08)	0.09 (0.08)	0.11 (0.07)	0.92 (0.49)	0.09 (0.04)	4.00	0.83	1.21	82		
CLE DET	6.15 (8.32)	–0.73 (0.50)	–0.07 (0.06)	–0.04 (0.06)	–0.12 (0.06)	0.16 (0.36)	0.05 (0.31)	1.42	0.36	1.33	90	55	8920
HTF NYC	–12.24 (10.17)	–0.29 (0.38)	–0.19 (0.05)	0.02 (0.05)	–0.04 (0.05)	0.88 (0.45)	0.09 (0.04)	2.32	0.93	1.42	106	51	10158
PHL DCA	–11.87 (4.99)	0.87 (0.57)	–0.13 (0.06)	0.03 (0.06)	–0.07 (0.06)	0.72 (0.21)	0.11 (0.03)	2.10	0.86	0.63	122	76	7275
PDX SEA	0.49 (1.79)	–0.07 (0.32)	–0.15 (0.05)	0.02 (0.05)	0.03 (0.05)	0.33 (0.09)	0.19 (0.02)	2.52	0.89	2.22	145	35	12329
CHI IND	–1.49 (2.27)	–0.76 (0.37)	–0.14 (0.04)	–0.04 (0.03)	–0.16 (0.03)	0.51 (0.11)	0.11 (0.01)	1.60	0.93	1.84	164	48	10603
BOS NYC	293.47 (30.14)	–1.31 (0.63)	–0.19 (0.07)	–0.03 (0.07)	–0.15 (0.07)	–11.33 (1.22)	–0.07 (0.07)	–9.10	0.87	0.80	188	3	85335
PIT DCA	–19.02 (4.32)	0.39 (0.49)	–0.15 (0.05)	–0.02 (0.05)	–0.16 (0.05)	1.09 (0.13)	0.11 (0.01)	3.10	0.94	1.11	191	82	6949
BAL NYC	–16.37 (7.34)	1.29 (0.54)	–0.17 (0.08)	0.03 (0.08)	–0.14 (0.08)	0.89 (0.31)	0.11 (0.03)	3.08	0.82	0.86	200	72	7481
DCA NYC	–22.98 (9.08)	–0.52 (0.46)	–0.20 (0.06)	0.03 (0.06)	–0.15 (0.05)	1.42 (0.36)	0.03 (0.03)	0.74	0.91	0.91	208	5	70980
DET PIT	–17.96 (9.25)	0.31 (0.41)	–0.15 (0.06)	–0.03 (0.06)	–0.15 (0.06)	1.04 (0.39)	0.07 (0.04)	1.42	0.83	1.98	208		
MIA TPA	–14.16 (8.99)	–0.29 (0.51)	0.23 (0.05)	–0.04 (0.05)	–0.23 (0.04)	1.13 (0.44)	–0.07 (0.09)	0.75	0.87	1.58	209		
DAL HST	–1.21 (0.73)	0.16 (0.21)	–0.09 (0.03)	–0.05 (0.03)	–0.11 (0.03)	0.40 (0.05)	0.13 (0.01)	1.91	0.97	1.25	225	29	14221
CHI DET	–14.23 (9.13)	0.05 (0.17)	–0.13 (0.03)	–0.02 (0.03)	–0.11 (0.03)	0.94 (0.39)	0.05 (0.04)	0.87	0.96	1.67	237	13	26816
DET IND	2.35 (1.11)	–0.01 (0.40)	–0.11 (0.04)	0.04 (0.04)	–0.11 (0.04)	0.17 (0.03)	0.31 (0.02)	2.72	0.86	2.37	240		
NYC ROC	–11.40 (1.74)	–0.45 (0.16)	–0.22 (0.02)	–0.03 (0.02)	–0.11 (0.02)	0.90 (0.08)	0.07 (0.01)	1.49	0.98	1.98	249	26	16795
CHI CIN	–9.26 (8.26)	–0.37 (0.30)	–0.02 (0.04)	–0.04 (0.04)	–0.08 (0.04)	0.79 (0.35)	0.07 (0.04)	1.13	0.93	1.69	252	61	8115
CIN PIT	–11.78 (8.33)	–0.09 (0.41)	–0.11 (0.06)	–0.00 (0.06)	–0.20 (0.06)	0.83 (0.33)	0.09 (0.03)	1.67	0.83	1.76	256		
PHL PIT	–19.56 (4.89)	0.08 (0.41)	–0.15 (0.04)	–0.03 (0.04)	–0.12 (0.04)	1.17 (0.17)	0.09 (0.01)	2.34	0.93	1.20	257	43	11596
CHI STL	–29.62 (6.38)	0.32 (0.25)	–0.19 (0.02)	–0.04 (0.02)	–0.12 (0.02)	1.66 (0.29)	0.03 (0.03)	0.75	0.98	1.32	261	23	16465
BOS PHL	–20.65 (8.20)	0.45 (0.37)	–0.20 (0.04)	–0.03 (0.04)	–0.18 (0.04)	1.15 (0.38)	0.09 (0.03)	2.50	0.93	0.92	270	31	13378
BUF NYC	–60.53 (14.39)	–1.12 (0.32)	–0.24 (0.03)	0.01 (0.03)	–0.05 (0.03)	3.06 (0.61)	0.01 (0.24)	0.43	0.95	1.71	291	11	27000
BAL CLE	–17.76 (4.64)	0.05 (0.16)	–0.19 (0.05)	–0.02 (0.05)	–0.21 (0.05)	1.12 (0.21)	0.09 (0.02)	2.29	0.93	1.78	305		
DCA CLE	–15.63 (4.48)	0.24 (0.41)	–0.18 (0.05)	–0.02 (0.05)	–0.22 (0.05)	0.98 (0.18)	0.07 (0.02)	1.45	0.93	1.98	305		
CHI CLE	–14.68 (6.48)	0.08 (0.28)	–0.15 (0.04)	–0.07 (0.04)	–0.14 (0.04)	0.96 (0.29)	0.07 (0.02)	1.46	0.94	2.80	307	20	16878
NYC PIT	–15.99 (4.66)	0.38 (0.40)	–0.24 (0.04)	–0.05 (0.04)	–0.11 (0.04)	0.93 (0.16)	0.11 (0.01)	3.35	0.91	1.36	315	8	31417
LAX SFO	–5.20 (17.04)	–0.39 (0.13)	–0.16 (0.06)	–0.12 (0.06)	–0.05 (0.06)	0.68 (0.74)	0.05 (0.09)	0.63	0.86	1.53	347	1	95875
CHI MSP	–21.34 (8.89)	–0.22 (0.29)	–0.15 (0.03)	–0.04 (0.03)	–0.11 (0.03)	1.32 (0.39)	0.05 (0.04)	1.20	0.97	1.44	354	16	19995
LAX PHO	–20.88 (6.41)	–0.05 (0.20)	–0.10 (0.03)	–0.08 (0.03)	–0.15 (0.03)	1.31 (0.30)	–0.01 (0.06)	0.16	0.96	1.99	356	30	13573

* STANDARD ERRORS IN PARENTHESES.

**DATA AVAILABLE FOR ROUTES RANKED FROM 1 TO 100 IN 1959.

TABLE 6 CONTINUED

CITY PAIR	CONSTANT	PRICE	SEA 1	SEA 2	SEA 3	γ	β	INCOME ELASTICITY	R^2	DURBIN WATSON	DISTANCE	PASSENGER RANK (1959)**	PASSENGERS (1959)**
CLE PHL	−27.69 (11.45)	−0.08 (0.25)	−0.21 (0.04)	−0.07 (0.04)	−0.19 (0.04)	1.53 (0.51)	0.05 (0.04)	1.34	0.95	2.50	358	89	6524
DEN SLC	−4.85 (2.89)	0.08 (0.40)	−0.09 (0.06)	−0.08 (0.06)	0.03 (0.06)	0.54 (0.09)	0.17 (0.01)	3.18	0.86	2.01	370		
BAL BOS	1.38 (6.74)	−1.98 (0.48)	−0.25 (0.07)	0.01 (0.07)	−0.12 (0.07)	0.65 (0.29)	0.11 (0.02)	1.94	0.90	0.49	393		
DCA BOS	−15.11 (8.09)	−1.51 (0.34)	−0.22 (0.04)	0.03 (0.04)	−0.12 (0.04)	1.30 (0.36)	0.03 (0.04)	0.63	0.96	0.75	393	24	16082
DET DCA	−14.49 (8.97)	0.18 (0.34)	−0.25 (0.05)	−0.01 (0.05)	−0.20 (0.05)	0.93 (0.41)	0.07 (0.05)	1.36	0.91	1.32	395	60	8123
CIN DCA	−11.57 (3.35)	−0.04 (0.07)	−0.17 (0.04)	−0.02 (0.04)	−0.10 (0.04)	0.81 (0.17)	0.11 (0.03)	2.37	0.96	1.08	403		
CLE NYC	−15.81 (4.66)	−0.14 (0.20)	−0.26 (0.02)	−0.04 (0.02)	−0.06 (0.02)	1.03 (0.20)	0.07 (0.01)	1.72	0.98	2.24	404	10	28801
CHI PIT	−11.66 (6.40)	−0.37 (0.55)	−0.23 (0.06)	−0.08 (0.06)	−0.18 (0.06)	0.87 (0.22)	0.11 (0.02)	2.70	0.86	1.85	409	46	11022
MKC MSP	−2.74 (1.35)	−0.49 (0.25)	−0.13 (0.03)	−0.02 (0.02)	−0.05 (0.02)	0.49 (0.07)	0.17 (0.01)	3.42	0.98	1.94	410		
CHI MKC	−8.76 (2.46)	0.42 (0.22)	−0.17 (0.03)	−0.03 (0.02)	−0.08 (0.03)	0.64 (0.12)	0.13 (0.01)	2.47	0.97	1.45	412	41	11867
DET PHL	−13.00 (12.83)	−0.14 (0.37)	−0.22 (0.05)	−0.07 (0.05)	−0.15 (0.05)	0.89 (0.53)	0.07 (0.05)	0.89	0.89	2.07	442	57	8339
DAL MKC	−6.62 (4.64)	−0.69 (0.41)	−0.35 (0.05)	−0.05 (0.05)	−0.15 (0.05)	0.75 (0.02)	0.11 (0.02)	3.36	0.93	2.34	452		
DET NYC	−27.72 (10.20)	−0.12 (0.25)	−0.29 (0.03)	0.02 (0.03)	−0.20 (0.03)	1.49 (0.41)	0.03 (0.03)	0.80	0.97	1.92	480	7	37663
INC DCA	0.62 (2.36)	−0.95 (0.43)	−0.26 (0.04)	0.04 (0.04)	−0.18 (0.04)	0.40 (0.04)	0.21 (0.01)	3.58	0.96	1.33	492		
CIN PHL	−0.58 (4.01)	−0.54 (0.65)	−0.16 (0.07)	−0.01 (0.07)	−0.11 (0.07)	0.33 (0.10)	0.27 (0.03)	4.85	0.85	2.10	501		
PDX SFO	−9.58 (3.11)	−0.45 (0.23)	−0.20 (0.03)	−0.13 (0.03)	−0.02 (0.03)	0.86 (0.14)	0.11 (0.02)	2.49	0.98	2.48	535	47	10761
BOS CLE	−31.44 (7.16)	0.05 (0.31)	−0.26 (0.03)	−0.05 (0.03)	−0.08 (0.03)	1.70 (0.32)	0.05 (0.02)	1.54	0.97	1.69	535		
CIN NYC	−14.68 (6.63)	−0.16 (0.27)	−0.20 (0.03)	0.03 (0.03)	−0.04 (0.03)	0.97 (0.26)	0.07 (0.02)	1.56	0.96	1.17	568	45	11463
LAX SLC	−3.97 (3.81)	0.14 (0.45)	−0.28 (0.05)	−0.14 (0.05)	0.02 (0.05)	0.45 (0.10)	0.19 (0.01)	3.45	0.91	1.84	579		
ATL CHI	−32.53 (0.33)	−0.34 (0.26)	−0.25 (0.03)	−0.06 (0.03)	−0.08 (0.03)	1.68 (0.41)	0.03 (0.05)	0.80	0.97	1.64	588		
BAL CHI	−2.11 (4.96)	−0.81 (0.37)	−0.27 (0.04)	−0.03 (0.04)	−0.16 (0.04)	0.58 (0.17)	0.13 (0.02)	2.39	0.95	0.69	595		
DCA CHI	8.70 (3.12)	−1.69 (0.34)	−0.21 (0.04)	0.01 (0.04)	−0.13 (0.04)	0.26 (0.09)	0.19 (0.01)	2.04	0.96	1.13	595	21	16764
SFO SLC	−3.20 (2.70)	−0.30 (0.33)	−0.21 (0.04)	−0.10 (0.04)	−0.00 (0.04)	0.53 (0.07)	0.15 (0.01)	2.55	0.95	1.85	598	97	6067
BOS DET	−16.69 (8.69)	−0.25 (0.48)	−0.32 (0.05)	−0.02 (0.05)	−0.02 (0.05)	1.06 (0.35)	0.09 (0.03)	2.35	0.94	0.83	612		
IND NYC	−9.12 (4.21)	−0.03 (0.40)	−0.29 (0.40)	0.01 (0.05)	−0.08 (0.05)	0.68 (0.05)	0.13 (0.12)	3.38	0.95	1.93	644	81	6949
NYC LOU	−6.31 (9.95)	−0.92 (0.20)	−0.25 (0.03)	0.05 (0.03)	−0.03 (0.03)	0.75 (0.43)	0.07 (0.03)	1.12	0.95	1.93	650	100	5908
CHI PHL	−6.81 (8.38)	−0.37 (0.37)	−0.27 (0.05)	−0.07 (0.05)	−0.17 (0.05)	0.65 (0.35)	0.11 (0.03)	2.06	0.90	2.29	664	28	14336
MIA ORL	−10.38 (5.37)	0.58 (0.52)	−0.08 (0.06)	−0.07 (0.06)	−0.21 (0.06)	0.69 (0.18)	0.11 (0.01)	2.20	0.86	1.95	668		
CHI NYC	−22.93 (14.43)	−0.62 (0.20)	−0.23 (0.02)	−0.02 (0.02)	−0.05 (0.02)	1.38 (0.57)	0.03 (0.04)	0.71	0.97	1.88	711	4	82624

*STANDARD ERRORS IN PARENTHESES.

**DATA AVAILABLE FOR ROUTES RANKED FROM 1 TO 100 IN 1959.

TABLE 6 CONTINUED

CITY PAIR	CONSTANT	PRICE	SEA 1	SEA 2	SEA 3	γ	β	INCOME ELASTICITY	R^2	DURBIN WATSON	DISTANCE	PASSENGER RANK (1959)**	PASSENGERS (1959)**
ATL NYC	−30.01 (6.68)	−0.21 (0.19)	−0.27 (0.02)	−0.02 (0.02)	−0.01 (0.02)	1.64 (0.28)	0.03 (0.03)	0.82	0.98	2.27	748	39	12029
CHI HTF	−9.31 (6.66)	0.00 (0.35)	−0.35 (0.04)	−0.11 (0.04)	−0.12 (0.04)	0.72 (0.27)	0.11 (0.03)	2.42	0.95	1.75	768		
CHI DAL	−5.73 (2.36)	−0.15 (0.26)	−0.18 (0.03)	−0.06 (0.03)	−0.09 (0.03)	0.59 (0.07)	0.11 (0.01)	1.95	0.97	1.39	803	95	6245
LAX PDX	−13.20 (6.99)	−0.63 (0.29)	−0.18 (0.03)	−0.11 (0.03)	0.03 (0.03)	1.02 (0.33)	0.07 (0.04)	1.42	0.97	2.19	826	85	6680
DEN LAX	−5.06 (4.56)	−0.29 (0.10)	−0.25 (0.03)	−0.07 (0.03)	0.12 (0.03)	0.61 (0.20)	0.11 (0.03)	1.76	0.97	1.39	830	50	10276
BOS CHI	−22.66 (8.19)	−0.65 (0.27)	−0.33 (0.29)	−0.08 (0.29)	−0.06 (0.03)	1.41 (0.34)	0.05 (0.03)	1.29	0.97	1.13	849	32	12898
NYC STL	−22.80 (6.57)	0.10 (0.25)	−0.29 (0.03)	0.05 (0.03)	0.03 (0.03)	1.24 (0.26)	0.07 (0.02)	2.02	0.96	1.92	873	27	14542
CHI DEN	−19.96 (9.32)	−0.23 (0.17)	−0.23 (0.03)	−0.08 (0.03)	0.05 (0.03)	1.22 (0.40)	0.09 (0.04)	2.89	0.98	1.91	917	93	6406
BAL MIA	−12.05 (5.47)	0.13 (0.37)	0.09 (0.05)	−0.07 (0.05)	−0.26 (0.05)	0.90 (0.25)	0.09 (0.23)	1.83	0.92	1.77	925		
DCA MIA	−0.54 (2.94)	0.01 (0.46)	0.04 (0.05)	−0.06 (0.05)	−0.25 (0.05)	0.36 (0.09)	0.17 (0.01)	2.47	0.88	1.31	925	52	9971
CLE TPA	−30.60 (18.49)	−0.05 (0.56)	0.26 (0.08)	0.02 (0.08)	−0.19 (0.08)	1.76 (0.97)	−0.01 (0.13)	0.27	0.77	1.88	930		
CHI HST	−17.02 (7.02)	−0.38 (0.23)	−0.21 (0.03)	−0.05 (0.03)	−0.11 (0.03)	1.11 (0.30)	0.05 (0.04)	1.05	0.98	1.50	941		
BAL MKC	−3.61 (1.26)	−0.12 (0.31)	−0.16 (0.03)	−0.06 (0.03)	−0.09 (0.03)	0.47 (0.04)	0.17 (0.01)	2.00	0.97	1.07	943		
MKC DCA	−26.70 (17.55)	4.44 (2.75)	−0.30 (0.21)	0.04 (0.21)	−0.05 (0.21)	0.37 (0.37)	0.31 (0.09)	5.43	0.21	0.51	943		
DEN SFO	−21.26 (7.63)	−0.38 (0.42)	−0.30 (0.05)	−0.10 (0.05)	0.13 (0.05)	1.36 (0.33)	0.05 (0.05)	1.19	0.93	1.66	947	98	6029
CIN MIA	−7.09 (4.35)	0.16 (0.50)	0.20 (0.05)	0.07 (0.05)	−0.08 (0.05)	0.61 (0.18)	0.11 (0.02)	1.95	0.88	2.54	954		
LAX SEA	−32.27 (13.84)	0.50 (0.46)	−0.20 (0.51)	−0.02 (0.05)	0.13 (0.05)	1.64 (0.66)	0.01 (0.11)	0.24	0.92	2.12	960	22	16619
CHI TPA	−5.70 (6.68)	−0.39 (0.26)	0.21 (0.04)	−0.00 (0.04)	−0.28 (0.04)	0.64 (0.33)	0.11 (0.04)	1.87	0.96	2.18	1000	90	6447
NYC TPA	−0.70 (5.02)	−0.06 (0.39)	0.13 (0.54)	−0.00 (0.05)	−0.18 (0.05)	0.37 (0.22)	0.13 (0.02)	1.72	0.80	1.89	1000	36	12249
MIA PIT	−12.68 (6.51)	0.22 (0.44)	0.22 (0.07)	0.04 (0.07)	−0.06 (0.07)	0.87 (0.30)	0.09 (0.03)	1.82	0.77	1.69	1013	79	7121
MSP NYC	−19.22 (8.44)	0.21 (0.35)	−0.27 (0.04)	0.11 (0.04)	0.07 (0.04)	1.07 (0.33)	0.07 (0.03)	1.78	0.94	1.71	1015	40	11913
DEN SEA	−21.94 (6.85)	1.11 (0.46)	−0.26 (0.07)	0.11 (0.07)	0.32 (0.07)	1.03 (0.31)	0.13 (0.05)	3.68	0.91	2.21	1019		
MIA PHL	−30.35 (14.82)	−0.11 (0.32)	0.24 (0.05)	−0.14 (0.05)	−0.26 (0.05)	1.74 (0.70)	0.03 (0.08)	0.75	0.94	1.70	1021	33	12501
MIA STL	2.80 (5.75)	−0.11 (0.80)	0.23 (0.08)	0.01 (0.08)	−0.16 (0.08)	0.19 (0.15)	0.19 (0.02)	1.84	0.67	1.29	1062		
MIA NYC	3.33 (4.05)	−0.49 (0.27)	0.25 (0.05)	−0.07 (0.05)	0.17 (0.05)	0.36 (0.17)	0.13 (0.02)	1.92	0.89	1.86	1094	2	95354
MKC NYC	−3.58 (2.67)	−0.31 (0.24)	−0.26 (0.03)	−0.06 (0.03)	0.03 (0.03)	0.50 (0.10)	0.15 (0.01)	3.16	0.96	2.36	1095	69	7629
CLE MIA	−0.89 (6.50)	−0.21 (0.41)	0.38 (0.07)	−0.02 (0.07)	−0.34 (0.07)	0.44 (0.29)	0.09 (0.03)	2.05	0.85	2.02	1098	59	8277
NYC ORL	−14.99 (6.96)	−0.17 (0.39)	−0.18 (0.03)	−0.04 (0.03)	−0.04 (0.03)	0.99 (0.23)	0.07 (0.01)	1.60	0.95	1.53	1170	71	7618
CHI MIA	−1.40 (5.74)	−0.64 (0.32)	0.30 (0.05)	−0.05 (0.05)	−0.32 (0.05)	0.56 (0.26)	0.11 (0.03)	1.89	0.93	1.50	1190	14	24732

* STANDARD ERRORS IN PARENTHESES.

** DATA AVAILABLE FOR ROUTES RANKED FROM 1 TO 100 IN 1959.

TABLE 6 CONTINUED

CITY PAIR	CONSTANT	PRICE	SEA 1	SEA 2	SEA 3	γ	β	INCOME ELASTICITY	R^2	DURBIN WATSON	DISTANCE	PASSENGER RANK (1959)**	PASSENGERS (1959)**
HST DCA	6.94 (6.20)	−2.21 (1.14)	−0.20 (0.12)	0.03 (0.11)	−0.14 (0.11)	0.37 (0.08)	0.27 (0.04)	5.36	0.80	0.87	1219		
DAL LAX	−0.16 (4.80)	−1.10 (0.53)	−0.24 (0.05)	−0.04 (0.05)	0.05 (0.05)	0.52 (0.11)	0.15 (0.13)	3.05	0.96	1.27	1238	42	11614
BOS MIA	−7.66 (3.86)	0.19 (0.23)	0.56 (0.05)	−0.12 (0.05)	−0.42 (0.05)	0.64 (0.19)	0.13 (0.02)	3.05	0.97	1.43	1257	42	11614
MKC LAX	−33.81 (8.03)	−0.20 (0.10)	−0.29 (0.02)	−0.01 (0.02)	0.14 (0.02)	1.83 (0.38)	−0.01 (0.05)	0.25	0.99	2.04	1353		
DAL NYC	−4.09 (3.41)	0.01 (0.31)	−0.20 (0.03)	0.06 (0.03)	0.00 (0.03)	0.48 (0.08)	0.13 (0.01)	2.52	0.97	1.63	1371	49	10509
HST LAX	−29.14 (10.20)	−0.17 (0.38)	−0.27 (0.03)	0.05 (0.03)	0.02 (0.03)	1.58 (0.37)	0.05 (0.03)	1.47	0.93	0.23	1371		
HST NYC	−4.74 (10.20)	−0.68 (0.38)	−0.22 (0.03)	0.01 (0.03)	−0.05 (0.03)	0.66 (0.37)	0.07 (0.03)	1.14	0.97	2.09	1419	65	7957
MIA MSP	−4.92 (4.17)	−0.37 (0.46)	0.61 (0.07)	−0.18 (0.07)	−0.60 (0.07)	0.57 (0.20)	0.15 (0.03)	3.50	0.95	1.78	1506		
LAX MSP	−9.07 (11.63)	−0.09 (0.39)	−0.26 (0.06)	−0.09 (0.06)	0.12 (0.06)	0.88 (0.50)	0.07 (0.06)	1.27	0.93	2.40	1524		
LAX STL	−9.18 (6.29)	−0.59 (0.31)	−0.29 (0.05)	0.03 (0.05)	0.15 (0.05)	0.79 (0.27)	0.11 (0.03)	2.46	0.95	1.40	1586		
DEN NYC	−12.49 (5.21)	−0.05 (0.33)	−0.28 (0.04)	0.10 (0.04)	0.28 (0.04)	0.81 (0.17)	0.13 (0.01)	3.88	0.97	1.65	1628	83	6853
CHI SEA	−72.13 (29.73)	0.38 (0.36)	−0.38 (0.07)	0.06 (0.07)	0.24 (0.07)	3.33 (1.31)	−0.03 (0.13)	1.23	0.91	1.17	1733		
CHI LAX	−1.95 (6.32)	−0.97 (0.10)	−0.16 (0.02)	−0.00 (0.02)	0.19 (0.02)	0.61 (0.28)	0.09 (0.03)	1.39	0.98	1.99	1742	12	26845
CHI SFO	−93.79 (11.71)	−0.94 (0.15)	−0.26 (0.02)	−0.01 (0.02)	0.14 (0.02)	4.44 (0.50)	−0.07 (0.04)	3.42	0.99	2.37	1858	25	15249
DET LAX	−6.22 (5.85)	−0.33 (0.39)	−0.25 (0.05)	0.00 (0.05)	0.18 (0.05)	0.61 (0.22)	0.13 (0.03)	2.58	0.95	1.40	1979	78	7178
CLE LAX	−23.07 (6.98)	0.15 (0.25)	−0.19 (0.04)	0.04 (0.04)	0.28 (0.04)	1.24 (0.28)	0.05 (0.03)	1.13	0.95	1.02	2045		
DET SFO	−19.71 (7.31)	−0.15 (0.24)	−0.28 (0.05)	0.02 (0.05)	0.15 (0.05)	1.14 (0.32)	0.09 (0.04)	2.47	0.96	0.92	2087		
CLE SFO	−30.00 (6.19)	0.01 (0.01)	−0.24 (0.04)	−0.01 (0.04)	0.18 (0.04)	1.58 (0.28)	0.05 (0.03)	1.44	0.96	2.07	2162		
BAL LAX	−9.56 (3.96)	0.26 (0.32)	−0.16 (0.04)	0.08 (0.04)	0.09 (0.04)	0.66 (0.14)	0.15 (0.02)	3.37	0.95	1.29	2296	64	8019
DCA LAX	−26.61 (4.57)	2.16 (0.76)	−0.10 (0.13)	0.16 (0.14)	0.15 (0.14)	0.85 (0.17)	0.21 (0.05)	7.60	0.88	0.82	2296		
LAX MIA	−13.79 (9.75)	−0.07 (0.60)	−0.13 (0.05)	−0.04 (0.05)	−0.11 (0.05)	0.89 (0.36)	0.11 (0.04)	2.90	0.96	1.94	2335		
LAX PHL	−17.22 (6.64)	0.42 (0.30)	−0.20 (0.05)	0.04 (0.05)	0.18 (0.05)	0.90 (0.27)	0.11 (0.03)	2.72	0.95	1.95	2389	99	5930
NYC SEA	−10.12 (5.09)	0.30 (0.46)	−0.26 (0.07)	0.18 (0.07)	0.30 (0.07)	0.60 (0.16)	0.17 (0.02)	4.52	0.92	1.51	2402		
BAL SFO	−13.62 (3.33)	0.06 (0.24)	−0.25 (0.32)	0.04 (0.03)	0.10 (0.03)	0.89 (0.03)	0.13 (0.14)	3.38	0.98	2.14	2436	77	7175
LAX NYC	−9.29 (8.07)	−0.67 (0.17)	−0.17 (0.03)	0.12 (0.03)	0.24 (0.03)	0.85 (0.33)	0.07 (0.03)	1.41	0.97	1.75	2446	6	48476
NYC SFO	−8.41 (5.06)	−0.93 (0.20)	−0.26 (0.03)	0.10 (0.03)	0.23 (0.03)	0.87 (0.21)	0.09 (0.02)	2.21	0.98	2.62	2566	9	29479
BOS LAX	−13.23 (6.56)	−0.26 (0.15)	−0.21 (0.04)	0.05 (0.04)	0.26 (0.04)	0.91 (0.28)	0.09 (0.03)	2.04	0.95	1.12	2591	67	7778
BOS SFO	−24.10 (6.06)	−0.15 (0.14)	−0.32 (0.04)	0.03 (0.04)	0.22 (0.04)	1.35 (0.28)	0.09 (0.03)	3.30	0.97	1.93	2693		

* STANDARD ERRORS IN PARENTHESES.

** DATA AVAILABLE FOR ROUTES RANKED FROM 1 TO 100 IN 1959.

fares. Since utilization of these fares varied depending on the type of travel on a route, the author computed indices for 37 separate routes where airline data on usage were available. These data were gathered for the CAB general passenger fare investigation.[19] Where utilization data were not available, weights for the route appearing to be most similar in distance and in travel characteristics were used.

Equation (14) was estimated for 115 city pairs using a special nonlinear regression program designed explicitly for the specification. For each city pair, a number of values of β were tried (normally on the interval $-0.3 < \beta < 0.3$). Then for each city pair the interval was progressively reduced around the particular β value that minimized the variance, and the interval of trial values for β decreased. The standard errors of the coefficients were calculated by inverting the Hessian at the likelihood function for (14) at the value of β which minimized the sum of the squared residuals. The results from the estimations are mixed. The income weighting coefficient, β, is almost always greater than zero and usually statistically different from zero regardless of the precise specification of the model. The price coefficient, η, on the other hand, does not demonstrate strong regularities.

By far the most satisfactory and strongest defense can be made for the conclusion that air travel demand is very income elastic. The results for the estimate of demand for all city pairs, given in Table 6, indicate that only 7 of the 115 β coefficients are less than zero and 79 of the remaining 110 coefficients are statistically greater than zero at a 95-percent confidence level. The second income coefficient, γ, demonstrates even more remarkable characteristics: only one estimate of γ is less than zero and 105 are statistically greater than zero at the 95-percent confidence level.

The second conclusion on the responsiveness of demand to price changes is relatively weak. No regular behavior for any group of city pairs could be observed. In terms of summary statistics, 72 estimates of price elasticity are statistically greater than -1.0 at the 95-percent confidence level while only 23 coefficients are statistically different from zero. The mean price elasticity is about -0.12 and the variance approximately 0.45. The wide variance indicates the inadvisability of interpreting aggregate measures of elasticities with any confidence. As one might expect, the response of travelers to changes in prices appears to be more predictable on vacation routes than business routes.

☐ **Estimates of the income coefficients γ and β.** We noted earlier that personal income was not necessarily the only motivating force in the demand for travel.[20] Indeed, a very persuasive argument could be made for the use of corporate income or cash flow; such a variable would reflect upon a corporation's propensity to travel. Unfortunately, while data on individual corporations might be useful in describing their expenditures on travel, it would not help to explain the distribution over cities of corporate trips. Thus it was necessary to choose some variable which measured the trip potential of each city.

---

[19] See [15].

[20] See p. 443 above.

TABLE 7

FREQUENCY TABLE FOR β IN 32–QUARTER ESTIMATION PERIODS

VALUE OF β	NO. OF OCCURENCES	MEAN VALUE OF γ
−0 07	3	−5.76
−0 03	1	3.33
−0.01	3	1.60
0 01	2	2.35
0.03	8	1.58
0.05	12	1.22
0.07	18	0.92
0.09	15	1.03
0.11	22	0.75
0 13	12	0.54
0.15	4	0.56
0 17	5	0.49
0 19	5	0.35
0.21	2	0.62
0.23	—	—
0.27	2	0 35
0.31	1	0.37

The results of this study indicate that the city-to-city air travel can be explained very well by incomes in the individual cities. These results are consistent in 108 of 115 cases with the survey data.[21] Furthermore, the coefficients almost always have the correct sign. The median income weighting coefficient, β, for all city pairs is between 0.09 and 0.11. The mean is approximately 0.11.[22]

Table 7 gives the distribution of β. The standard error of the distribution (0.04) suggests that there are large differences in the income coefficients for different routes. Since the individual β's possess very small variances, the large variance of the mean suggests further biases in aggregative approaches. The second mass coefficient, γ, does not appear to be distributed independently of the β_i. Table 7 also tabulates the average value of γ by the value of β. Aside from the outliers for $\beta = -0.07$, the regularity of the decline seems fairly systematic. This result indicates that as the influence of higher incomes on the size of the mass variable increases, the influence of the mass variable on traffic decreases. The net effect appears to lead to an increase in travel.

The estimated values of β and γ do not indicate whether a city pair would be more "income responsive" and generate more traffic with a large β than a small β, since such an effect would depend not only on the value of β, but also on γ and the effect of a change in β on the size of M. There are two ways of calculating the effect of such a change. The first involves the "calculation" of income elasticity.

Although an estimate of income elasticity is not given directly by the model, it is possible to calculate one.[23] The calculated income elasticity for each route is given in Table 6. Although these

---

[21] Surveys obtained by on-board passenger questionnaires as well as those developed from door-to-door interviewing by many different research organizations have supported the assumption that trips increase with income. While the numbers vary, the indications are that if a family with income less than $5,000 per year takes one person air trip per year (one air trip by one person), then the family with income in the $10,000 range will take between 1½ and 2 person air trips, the family with income in the $10,000 to $15,000 will take from 4½ to 6 person air trips, and the family with income above $15,000 will take from 16 to 24 trips. As an example, the following table tabulates the results from the 1967 U. S. Census of Transportation (see [13], p. 20 and [14], p. 23)

Income Range (000 omitted)	Percent of Respondents	÷	Percent of Household In Income Group	=	Air Trips Per Household
0–2	1.4		13 0		0 11
2–3	1.3		7 2		0.18
3–4	2.3		6.9		0.33
4–5	2.2		6.6		0.33
5–6	4.5		7.6		0.59
6–7.5	8.2		12.2		0.65
7.5–10	14.6		16.6		0.88
10–15	27.4		19.1		1.43
15–25	18.8		8.1		2.32
25 + above	9.4		2.1		4 48.

[22] A coefficient of $\beta = 0.11$ implies that households with incomes above $20,000 take roughly 16 times as many trips as do households with average incomes below $10,000. A precise estimate of the relative number of trips cannot be given since the ratio depends upon the average income $\bar{Y}_i$ for those households with incomes below $5,000 in city i and the average income $\bar{Y}_i^{20}$ for those households with income above $20,000

[23] A direct computation of an income elasticity is not possible. However, through

income elasticities are generally greater than unity and thus quite satisfactory, they must be treated with some skepticism since they involve the product of γ and β. A larger value of β, however, generally coincides with a larger estimated income elasticity, as can be observed from Table 8, where the estimated income elasticities are cross-classified by the values of β for individual city pairs.

The second method of determining income responsiveness involves the use of income distributional data on X^K and $\overline{Y}^K$ and various average estimates of $\overline{\gamma}$ to calculate a net effect. This method involves first calculating mass variables, M, for various values of β and then examining the values of M^γ. In such a test mass can be written as a function of β alone if the income distribution $f(x)$ and average income $\overline{Y}^K$ remain unchanged:

$$M(\beta) = \sum_{K=1}^{N} X^K e^{\beta \overline{Y}^K}. \tag{16}$$

One can then calculate the values of γ^* corresponding to $M(\beta)$ such that the term $Z = M(\beta)^\gamma$ remains constant for all values of β. In Table 6 the calculated values of γ^* are tabulated. For this analysis, aggregate income distribution data for the entire United States,

TABLE 8

INCOME ELASTICITIES CLASSIFIED BY VALUE OF β FOR 32–QUARTER SAMPLE

β	INCOME ELASTICITY
−0.07	−2 66
−0 03	1 23
−0.01	0.22
0.01	0.33
0.03	0.75
0 05	1 21
0.07	1.43
0 09	2.32
0 11	2 39
0.13	2.90
0.15	3 13
0.17	3 12
0 19	2.44
0.21	5.59
0 27	5.10
0.31	4.10

---

the following analysis, one can be obtained. The essential part of the demand equation is given by

$$T_{ij} = aP^{\eta}(M_i M_j)^{\gamma}, \tag{i}$$

where

$$M_i = \sum_{K=1}^{N} X_i^K e^{\beta \overline{Y}_i^K}$$

An estimate of $(\delta T_{ij}/\delta \overline{Y})\ (\overline{Y}/T_{ij})$ is desired. Suppose any Y_i^K can be given by $\alpha_K \overline{Y}$. Income elasticity η_Y can be written as

$$\begin{aligned} \eta_Y &= \frac{\delta T_{ij}}{\delta(M_i M_j)} \frac{\delta(M_i M_j)}{\delta \overline{Y}} \frac{M_i M_j}{T_{ij}} \frac{\overline{Y}}{M_i M_j} \\ &= \gamma \frac{\partial M_i M_j}{\delta \overline{Y}} \frac{\overline{Y}}{M_i M_j} \end{aligned} \tag{ii}$$

Due to the assumption $Y_i^K = \alpha_K \overline{Y}$, $\delta(M_i M_j)/\delta \overline{Y}$ *is given by*

$$\begin{aligned} \frac{\delta(M_i M_j)}{\delta \overline{Y}} = &\sum_{K=1}^{N} X_i^K e^{\beta \overline{\alpha_K} Y} \sum_{K=1}^{N} X_j^K \beta \alpha_K e^{\beta \alpha_K \overline{Y}} \\ &+ \sum_{K=1}^{N} X_i^K \beta \alpha_K e^{\beta \alpha_K \overline{Y}} \sum_{K=1}^{N} X_j^K e^{\beta \alpha_K \overline{Y}} \end{aligned} \tag{iii}$$

By substituting $Y_i^K/\overline{Y}$ for α_K, (iii) can be written as

$$\frac{\delta(M_i M_j)}{\delta \overline{Y}} = \frac{\beta}{\overline{Y}} (M_i \hat{M}_j + \hat{M}_i M_j), \tag{iv}$$

where

$$\hat{M}_j = \sum_{K=1}^{N} X_i^K \overline{Y}_i^K e^{\beta \alpha_K \overline{Y}}.$$

Thus η_Y can be written as

$$\eta_Y = \hat{\gamma} \beta \left\{ \frac{M_i \hat{M}_j + \hat{M}_i M_j}{M_i M_j} \right\}, \tag{v}$$

where $\hat{\gamma}$ and $\hat{\beta}$ are estimates of γ and β.

as given in the *Statistics of Income for 1967*,[24] were used for X_l and Y_l. Along with the values of γ^*, the values of $M(\beta)^{\bar{\gamma}(\beta)}$,

$$M(\beta)^{\bar{\gamma}(\beta)} = \left[\sum_{K=1}^{N} X^K e^{\beta \bar{Y}^K}\right]^{\bar{\gamma}(\beta)}, \tag{17}$$

are given. (The value of $\gamma(\beta)$ was taken from Table 7.) This term can be called the average aggregate value of mass for a value of β. If the approach used in this paper is neutral, then increases in β will be offset by decreases in γ. This value of γ (which totally offsets the increase in β) γ^*, can be calculated by the formula

$$\gamma^*(\delta) = \frac{(\gamma(\beta)^* \log(M(\beta))}{\log(M(\delta))}, \tag{18}$$

since neutrality implies

$$M(\beta)^{\bar{\gamma}(\beta)} = M(\delta)^{\bar{\gamma}(\delta)} \tag{19}$$

for any value of δ.

In Table 9 we have tabulated the values of γ^* and $\bar{\gamma}$ assuming $\delta = 0.01$. The neutrality assumption on the offsetting effects of γ and β appears to be rejected. This implies that a shift in the income distribution to upper incomes that leaves average income unchanged will cause an increase in travel.

TABLE 9

VALUES OF THE MASS VARIABLES FOR VALUES OF β AND $\bar{\gamma}$ (β)

β	M (β)	M $(\beta)^{\bar{\gamma}(\beta)}$	$\bar{\gamma}$	γ^*
−0.07	0.894	1.550	−5.76	−0.55
−0.05	0.915	0.825	—	−0.68
−0.03	0.943	0.822	3.33	−1.09
−0.01	0.877	0.811	1.60	−0.44
0.01	1.024	1.057	2.35	2.35
0.03	1.085	1.112	1.58	1.42
0.05	1.168	1.209	1.22	0.28
0.07	1.286	1.260	0.92	0.22
0.09	1.448	1.464	1.03	0.15
0.11	1.703	1.492	0.75	0.10
0.13	2.062	1.478	0.54	0.08
0.15	2.631	1.719	0.56	0.06
0.17	3.499	1.847	0.49	0.04
0.19	4.996	1.757	0.35	0.03
0.21	7.020	3.348	0.62	0.03
0.23	10.463	2.202	0.30	0.02
0.25	16.451	2.317	—	0.02
0.27	24.199	2.624	0.35	0.02
0.29	38.560	2.991	0.35	0.02
0.31	61.530	7.277	0.37	0.01

γ^* IS DEFINED SO THAT M $(\beta)^{\gamma^*}$ = M $(0.01)^{\gamma\,(0.01)}$

☐ **Estimates of the price coefficient.** The poor estimates of the price coefficients appear to offset the very strong estimates of the income coefficients. Many of the price coefficients are greater than zero and more than half are not statistically different from zero. Perhaps such results should have been expected since there were only two general price changes that affected all routes (although there were other price changes on each route as jets, coach service, and discount fares were introduced) and, on some of the major transcontinental routes, there was some price competition in both first class and coach. However, many routes received infrequent price changes. Income, on the other hand, increased throughout the period both in aggregate and per capita terms. Yet the price changes were sufficiently large that they should have had measurable results if airline traffic were truly price sensitive.

---

[24] [17].

When the routes are arranged in order of annual origin-destination traffic and the coefficients examined in this new order, it becomes apparent that the price coefficient can be generally estimated with greater accuracy on the major routes. The results indicate that 11 (52 percent) of the price coefficients are statistically less than zero at the 95-percent confidence level for routes among the 25 most traveled in 1959. While some accuracy was gained in terms of reduced standard error, they do not tend to be statistically different from -1, which means that one can make no conclusions with respect to the price elasticity of demand.

For the routes ranked from 26 to 100 in terms of 1959 traffic density, 38 percent of the price coefficients are greater than zero and only 11 percent are significantly less than zero. This is a much lower rate of success than that demonstrated by the top 25 routes. On the other hand, 71 percent of the estimated price elasticities are significantly greater than -1 at the 95-percent confidence level. The other 43 routes which are not included in either group demonstrate even poorer results, as 40 percent of the price coefficients are greater than zero. This suggests that the variance of the price coefficient decreases as traffic (and presumably service) increases. In cross-sectional models, one should presumably adjust for this bias by weighting observations.

When the point estimates of the price elasticity for these two groups are classified for broad distance categories, a clearer pattern begins to emerge. In Table 10 these coefficients are cross-classified by four distance categories, three coefficients intervals, and market size. The estimates for routes of less than 300 miles are fairly evenly distributed between estimates that are greater than zero, less than zero but greater than -1, and less than -1. However, as distance increases, most of the estimates fall in the middle range, between zero and -1. We observe then that price elasticity becomes less random and tends towards inelasticity as distance increases, and that the error of estimate (which can be observed by comparing parts (a) and (b) of Table 10) decreases as market size increases.

Further classification of these results can be performed by analyzing the coefficients obtained for certain types of routes. For cities that are within an easy day's drive of one another, the point estimates of the price elasticity are usually very close to zero. For cities that are more than a day's drive but less than 1000 miles apart, the estimates seem to fall around -0.6. For cities farther apart, the estimates seem to be fairly close to -1.0. There are exceptions, however, to the first two groups of cities.

In summary, it would appear first that the shorter, predominantly business routes demonstrate very small price elasticities. It seems further apparent that the later introduction of jets and other innovations such as the Eastern Shuttle caused an outward

TABLE 10

(a) CLASSIFICATION OF PRICE COEFFICIENTS FOR THE TOP 100 ROUTES BY DISTANCE: ESTIMATION PERIOD 1960–1967

DISTANCE INTERVAL	POINT ESTIMATE OF η, THE PRICE COEFFICIENT		
	$\eta > 0$	$0 > \eta > -1$	$-1 > \eta$
$d < 300$	2* 7**	1 5	2 0
$300 \leq d < 600$	2 2	4 8	2 0
$600 \leq d < 100$	1 2	1 10	0 0
$d \geq 1000$	0 8	6 11	0 0
ALL DISTANCES	5 19	12 34	4 0
(b) PRICE COEFFICIENTS SIGNIFICANTLY DIFFERENT FROM ZERO CLASSIFIED BY DISTANCE			
$d < 300$	0 0	0 2	2 0
$300 \leq d < 600$	0 0	1 1	2 0
$600 \leq d < 1000$	0 0	1 4	0 0
$d \geq 1000$	0 0	5 0	0 0
ALL DISTANCES	0 0	7 7	4 0

* NUMBER IN UPPER CORNER OF SQUARE REPRESENTS THE NUMBER OF PRICE COEFFICIENTS FROM TOP 25 ROUTES IN THIS GROUP.

** NUMBER IN LOWER CORNER OF SQUARE REPRESENTS THE NUMBER OF PRICE COEFFICIENTS FROM ROUTES RANKED FROM 26 TO 100 IN THIS GROUP.

shift in demand which probably affected the estimates of coefficients on shorter routes. Second, price elasticity increases in absolute value with distance. Third, price elasticity apparently approaches a value of unity in the limit.

6. Conclusion

■ The results presented in this paper indicate that parameters estimated in aggregate, or large-scale cross-section models of the demand for air transportation, are, at best, the average of micro parameters and possess very large variances. Further, given the analysis of Theil,[25] there is no reason to believe that they are unbiased averages. The analysis also demonstrates that gravity models which use population, per capita income, or combinations of the two fail to take into account the significant impact of changes in income distribution.

The results should not, however, be interpreted as final estimates of the demand for air travel between any two cities. The intention of the paper is to show that there are large differences in the demand for air travel that can not be ignored or overlooked as they are in aggregate or cross-sectional models. The failure to take account of differences in demand due to regional fluctuations, or, more significantly, differences in the trend of airline service to individual cities only serves to accentuate the impossibility of accepting the results of aggregate or cross-sectional analyses of demand. The analysis of such differences and thus the refinement of point-to-point models require further study.

---

[25] In [12]

References

1. Bartlett, H. C. "The Demand for Air Transportation, 1947–1962." Unpublished Ph.D. dissertation. University of Michigan, 1965
2. Brown, S. and Watkins, W. S. "The Demand for Air Travel: A Regression Study of Time Series and Cross-Sectional Data in the U. S. Domestic Market." *Highway Research Record*, No. 213 (1968), pp. 21–34.
3. Chow, G. "Tests of Equality Between Subsets of Coefficients in Two Linear Regressions." *Econometrica*, Vol. 28, No. 3 (July 1960), pp. 591–605.
4. Clemins, E. "Price Discrimination and the Multi-Product Firm," in Heffelbower and Stocking, eds., *Readings in Industrial Organization and Public Policy*, Homewood, Ill. · Richard D. Irwin, Inc., 1958.
5. Gronau, R. "The Value of Time in Passenger Transportation: The Demand for Air Travel." National Bureau of Economic Research, Occasional Paper 109. New York: Columbia Univ. Press, 1970.
6. Institute for Defense Analysis. "Demand Analysis for Air Travel by Supersonic Transport." Institute for Defense Analysis, Report R-118, Commerce Clearing House, Washington, D C., 1966.
7. Joun, Y. P. "The Demand for Air Travel." Unpublished Ph.D. dissertation. University of Washington, 1966.
8. Mathematica. *Studies in Travel Demand.* Vols. 1–4. Princeton, N.J.: Mathematica, 1967–1969.
9. Quandt, R. E "Some Perspectives on Gravity Models," in *Studies in Travel Demand* [8], Vol. 1.
10. ——— and Baumol, W. J. "The Demand for Abstract Transport Models, Theory and Measurement." *Journal of Regional Science*, Vol. 6, No. 2 (Winter 1966), pp. 13–26.
11. Systems Analysis and Research Corporation *Demand for Intercity Travel in the Boston-Washington Corridor.* Boston: Systems Analysis and Research Corp., 1964.
12. Theil, H. *Linear Aggregation of Economic Relations.* Amsterdam: North-Holland Publishing Co., 1954.
13. U. S. Bureau of the Census "Income in 1968 of Families and Persons in the U.S.A." *Current Population Reports*, Series P-60, No. 66. Washington, D.C.: U.S. Government Printing Office, 1969.
14. ——— *National Travel Survey, 1967.* Vol. 1. Washington, D.C.: U.S. Government Printing Office, 1969.
15. U. S. Civil Aeronautics Board. *Domestic Passenger Fare Investigation.* Docket 21866–5, 1970.
16. ——— *Origin-Destination Survey of Domestic Airline Passenger Traffic.* Washington, D. C. Air Transportation Association, quarterly 1960–1967 issues.
17. U. S Internal Revenue Service, Department of the Treasury. *Statistics of Income · Individual Income Tax Returns.* Washington, D. C.: U. S. Government Printing Office, 1959–1967.
18. Verleger, P. K. and Eads, G. "Statistical Biases in Aggregate Time Series Estimates of the Demand for Air Travel," in *1970 Proceedings of the Business and Economic Statistic Section of the American Statistical Association*, Washington, D. C.: May, 1971.

Some aspects of the pure theory of corporate finance: bankruptcies and take-overs

Joseph E. Stiglitz
Professor of Economics
The Cowles Foundation
Yale University

This paper considers the implications of bankruptcies, take-overs, and divergent expectations for the financial policy of the firm; we argue that, under reasonable assumptions, there is an optimal debt-equity ratio. Previous studies have shown that under very general conditions, if there is no chance of bankruptcy, then financial policy has no effect on the value of the firm; there is no optimal debt-equity ratio. Under certain very restrictive conditions, the no bankruptcy condition may be removed. We show that when these restrictive conditions are not satisfied, and when there is a real possibility of bankruptcy if the firm issues too much debt, the firm's valuation will depend on its debt-equity ratio; the real decisions of the firm (e.g., its investment and choice of technique) cannot be separated from its financial decisions; and the real decisions of the firm may not be productively efficient. Finally, the implications of the possibility of a take-over for the financial policy of the firm are considered.

1. Introduction

■ This paper argues that, under reasonable assumptions, there is an optimal debt-equity ratio. Previous studies[1] have shown that, under very general conditions, if there is no chance of bankruptcy, then financial policy has no effect on the value of the firm: there is no optimal debt-equity ratio. Under certain very restrictive conditions, the no-bankruptcy condition may be removed. But what happens when these restrictive conditions are not satisfied, when there is a real possibility of bankruptcy, if the firm issues too much debt? We shall show that the firm's valuation will depend on its debt-equity ratio (Section 3); that there will be, as a consequence, an optimal debt-equity ratio (Section 4); that the real decisions of the firm (e.g., its investment and choice of technique) cannot be separated from its financial decisions (the two must be made simultaneously); and that the real decisions of the firm may not be productively efficient.

Joseph E Stiglitz received his Ph.D degree from M.I.T. He is Professor of Economics at Yale University. Professor Stiglitz is coeditor of the *Journal of Public Finance* and associate editor of the *Journal of Economic Theory* and the *Review of Economic Studies.*

This paper is a revised version of a lecture delivered at a conference at Hakone, June 25–26, 1970. The author is very much indebted to H. Uzawa, who organized the conference, and to the other participants for their helpful comments and discussion. He is also indebted to R Merton for his comments on an earlier draft. The research described here was supported by the Guggenheim Foundation, the National Science Foundation, and the Ford Foundation.

[1] See Modigliani and Miller [14] and Stiglitz [21, 24].

In addition to its treatment of bankruptcy, the detailed example considered in this paper differs in one other respect from most other recent models[2] of capital markets: it explicitly takes into account differences in expectations. In our view, the assumption of identical expectations is unrealistic, some of its implications are untenable,[3] and it leaves unexplained some important phenomena of the capital market. One of these, take-overs, will be discussed in some detail in Section 5.

The reason that the assumption of identical expectations has been so extensively employed is partly that it is apparently difficult to introduce heterogeneous expectations into such models and still obtain simple results,[4] partly a feeling that one can explain too much too easily in terms of differing expectations, and partly that many of the phenomena which differing expectations might be used to explain can be explained in other terms. For instance, differences in portfolios held by different individuals can be explained by differences both in attitudes towards risk and in expectations. Both elements are difficult to observe directly; thus it is difficult to ascertain the relative importance of each.

Earlier literature tended to focus on differences in expectations,[5] more recent literature on differences in attitudes towards risk. The reason that different models focus on one or the other is not that either is the "true explanation"; rather, to understand how each of them works they are best studied in isolation—for example, to understand the effects of differences in attitudes towards risk on portfolio allocation we assume that individuals have the same expectations, and conversely. In the example which we shall present later, we assume that all individuals have the same risk attitude, and for simplicity we assume risk neutrality.

2. Bankruptcy and the independence of financial and real decisions

■ In this section, we attempt to define what we mean by bankruptcy and to suggest why it plays an important role in the analysis of firm behavior.

□ **Definition of bankruptcy.** In a two-period model in which the firm invests in the first period and makes its return in the second period (and thereupon dissolves), bankruptcy is easy to define: the income of the firm is less than the fixed obligations to bondholders.

---

[2] In particular, those growing out of the Sharpe-Lintner model. See [13], [15], and [18].

[3] Not only do the mean-variance models with identical expectations imply the irrelevance of financial policy, but they imply that the reason that financial policy is irrelevant is that there is a single mutual fund containing all corporate bonds and stocks. As in the Arrow-Debreu model (see [2] and [4]), the only risks that would not be diversified out would be risks such as those associated with the business cycle. The value of a firm formed by the merger of two other firms would be just the sum of the values of the constituent firms. For a more extended treatment of these points, see [23].

[4] Cf. Lintner [12].

[5] It is interesting to note that both Keynes, in explaining the liquidity preference schedule, and Knight [11], in explaining the continuation of entrepreneurship in spite of its apparent zero return, placed heavy emphasis on the role of differing expectations. See also Tobin [28].

In an ongoing firm, however, income can be less than the obligations of the firm and yet the firm is not bankrupt: it can simply borrow more. Clearly, a firm is bankrupt if the value of its equity is zero (it cannot be negative under conditions of limited liability), or equivalently, if the value of its future income streams, assuming it does not go bankrupt,[6] is less than the value of its outstanding debt. It is this definition which I used in my proof of the irrelevance of financial policy in a multi-period model.[7] (The argument required that no individual believed that in any state of nature at any date the value of equity would be zero.) Two questions are raised by this definition. First, should the debt be evaluated at market prices or at maturity values? One can imagine a situation where market prices exceeded maturity values and the firm was able to meet its debt obligations in the next period when they became due. Is it possible that the value of debt at market prices equals or exceeds the value of the firm in this period, and hence the value of the equity of the firm is zero? The answer is no; since in the next period, in all states of nature, there will be a residual left-over for the shareholders, the value of the equity in this period must be positive, and hence the value of the firm in this period must exceed the value of bonds (at market prices) in this period.[8] Second, although this is a sufficient condition for bankruptcy, is this a necessary condition? An investment plan in which there is a possibility of bankruptcy necessitates some individual believing that at some date in the future and in some state of nature the value of the firm would be less than the value of its bonds (valued as if they were "safe bonds") were the firm not to go bankrupt today, but at the actual date of formal bankruptcy the value of the equity of the firm may be positive. The reason for this may be demonstrated by example.[9]

Let the firm have a single asset, cash, say \$975,000. The firm has outstanding \$1,050,000 in bonds, each promising to pay one dollar a period from now. The interest rate is 10 percent, so the present value of the bonds is \$953,637, which is less than the value of cash. The firm has a single project; it pays \$1,150,000 with probability 0.25 and \$960,000 with probability 0.75. There is thus a large probability that the firm will not be able to meet its debt obligations. Yet the value of the equity in this period is positive; if the market were risk neutral, the value of the equity would be $(\$100{,}000 \times 0.25)/1.1 = \$22{,}727$. The expected return to the bondholders, however is not \$1,050,000 but only $\$1{,}050{,}000 \times 0.25 + \$960{,}000 \times 0.75 = \$982{,}500$ with a present value of $(\$982{,}500)/1.1 = \$893{,}181$, which is much less than the value of the bonds today if they were redeemed in this period (i.e., \$953,637). One might imagine a debt contract in which, if the bondholders could show that the policies of the firm would with significant probability result in a default of their bonds in the near future, they could declare the firm "bankrupt" in this period. In that case, the above firm could be declared bankrupt in this period even though after bankruptcy the value of the equity is positive (\$975,000 – \$953,637 = \$21,363).

---

[6] If there are Arrow-Debreu securities, we can easily calculate the value of future income streams; however, if there are not this may be hard to ascertain

[7] See Stiglitz [21]

[8] In [21] we show that if the value of equity is not zero at any date in any state of nature then the value of equity plus the market value of bonds of the firm is independent of the financial policy of the firm.

[9] I am indebted to Robert Merton, both for originally raising this question and for providing this example.

Bondholders may not be able to make the firms declare bankruptcy unless the firm has defaulted on their obligations to them; that is, even though there is a significant prospect of the firm's not being able to meet its debt obligations in the future and indeed, even though the value of the assets of the firm today are less than the present value of its outstanding debt, the bondholders cannot force bankruptcy today. For instance, as a slight modification of the above example, assume that the firm had only \$800,000 cash but investment opportunities still yielded \$1,050,000 with probability 0.25 and \$960,000 with probability 0.75. The present value of debt obligation (\$953,637) is greater than the cash (the capital of the firm), so one might say that the firm has negative net worth, i.e., is "bankrupt". Indeed, one might even observe that if the market were risk neutral, the present value of the entire income of the firm

$$\frac{\$960{,}000 \times 0.75 + \$1{,}050{,}000 \times 0.25}{1.1} = \$915{,}908$$

is less than the present value of the outstanding debt obligations. Yet the value of the equity is the same as in the previous example, \$22,727, so long as the bondholders cannot force bankruptcy. (If they can, clearly the value of the equity in this example is zero.)

These examples satisfy our condition for bankruptcy: at some date at some state of nature the firm is unable to meet its debt obligations; the value of equity is zero. But they also make clear that our definition does not resolve the timing of bankruptcy, i.e., whether the possibility of default in the future can allow the bondholders to force a bankruptcy now. Presumably, this is a matter of how the debt contract is written. Note that this problem is not so serious with short-term bonds as it is with long-term debt. With short-term bonds the individual can decide in each period whether, in effect, to demand his cash back. With long-term bonds there may be a change in the management of the firm after the individual has purchased the bond: the new management may undertake a policy which in the judgment of the bondholder—and perhaps of the market as a whole—will lead to bankruptcy with a high probability. Yet there may be nothing the bondholder can do.[10]

These problems are exacerbated by the clear conflict of interest between the bondholders and the equity owners. Because of the limited liability provided by bankruptcy, the firm is concerned only with the return in those states in which it does not go bankrupt.[11]

---

[10] There are, of course, certain legal constraints on what the firm can do

[11] Before the bonds are issued, the firm must clearly worry about returns in all states of nature, since that will affect the terms at which it can borrow, once having obtained the loans, however, the firm is only concerned about returns in those states in which the firm does not go bankrupt. (Of course, a firm which must have continual recourse to the capital market must continue to worry about returns in all states.)

For instance, in both of the examples presented above it was in the interest of the stockholders for the firm to undertake the project. If it does not, they receive in the first example $21,363, in the second nothing; if it does, they receive an amount whose present value (assuming risk neutrality) we have calculated at $22,727. In both cases the bondholders receive more if the firm dissolves in the first period, even though in one of the examples bondholders receive less than the present value of the debt obligations (but this is still larger than the present value of the expected return if the firm undertakes the project). Notice that in the first example, where the firm had $975,000 initial cash position, the expected real return to the investment was negative, (−8 percent) and yet, if the stockholders were making the decision, the project would be undertaken. It is clear that this may result in productive inefficiency.[12]

☐ **Implications of bankruptcy for financial decisions of the firm.** Bankruptcy is important in the financial decisions of the firm in three respects. First, no bankruptcy means that the nominal rate of interest which the firm must pay on its bonds is independent of its debt-equity ratio. If there is a chance of bankruptcy, bonds become risky assets; the nominal rate of interest[13] will rise as the firm borrows more. There is no reason to suppose[14] that the nominal rate should be the same function of the debt-equity ratio for all firms or individuals; indeed, quite the contrary. A complete theory must provide an explanation of the determination of the nominal rate for any given firm. We shall attempt to do this below.

One feature of the bond which is important in determining the optimal debt-equity ratio should be noted here: the bond is a promise to pay a fixed amount provided the firm does not go bankrupt, and to pay a pro rata share of the assets of the firm if it does. The borrower and the lender may, and in general will, have different attitudes about the mean and variance of the return to a risky bond. Borrowers usually will be more confident of repaying. From their view, the mean amount they are paying is larger than the mean amount the lender believes he is receiving (and the variance is smaller); the divergence between the two increases as the probability of bankruptcy, in the judgment of the lender, increases, (as the debt-equity ratio increases). In other words, the nominal interest rate which the borrower has to pay rises faster as the debt-equity ratio increases than the borrower feels is "justified."

Bankruptcy is also important to the financial decisions of the firm in that, as we have shown,[15] without bankruptcy the consumption opportunity sets of the individual are unaffected by the debt-equity ratio; with bankruptcy, except under special circumstances, this is no longer true.

---

[12] When there is a finite probability of bankruptcy, the rule of firm value maximization is not equivalent to maximizing the value of equity, and it is clearly the latter with which firms are concerned.

[13] The nominal rate of interest is that rate which the firm pays in all those states of nature in which it does not go bankrupt.

[14] As Modigliani and Miller seemed to have done in [14]. As we shall note below, their result, that the debt-equity ratio does not affect the value of the firm even with bankruptcy, depends crucially on this assumption.

[15] In [21, 24].

Consider, for example, the standard two-period model in which the firm invests in one period and makes its returns in the second. Consider two alternative financial policies—in one the firm issues B bonds, in the other no bonds (and hence there is no chance of default). It is easy to see that the individual buying stock on margin in the latter case can exactly replicate (if the value of the firm remains unchanged) the returns in the former case in those states of nature where the firm does not go bankrupt. But if the firm goes bankrupt in some state, θ', in the one case his return is zero, while in the other his return (after repaying his debt) is negative. In other words, the firm which issues a risky bond is issuing two securities. When the firm does not go bankrupt it is only issuing one security not already provided on the market; the bonds it issues are just like those of any other firm. Bankruptcy may, by creating a new security, change the consumption opportunity set of society.

The conditions, then, under which the possibility of bankruptcy does not affect the market valuation of a firm are of two kinds:

(1) Where bankruptcy does not in fact result in a new security, and

(2) Where the creation of the new security by bankruptcy makes no difference.

Condition (1) will occur under the following conditions:[16]

(1a) If there already exist as many securities as states of nature (the Arrow-Debreu model, which has been discussed elsewhere and shown to be "unrealistic").[17]

(1b) If the particular securities created by bankruptcy could have been created by individuals. This will be the case if individuals can purchase securities on margin[18] and create limited liability arrangements whereby (again in our two-period model) if profits of the firm are less than the nominal payments due on the loan, the shares are defaulted without further consequences to the individual.

It should be clear, however, that the scope for such limited liability arrangements is limited, and thus the presumption that bankruptcy does result in a new security remains.[19]

There is only one general situation in which the new securities created by bankruptcy make no difference; that is, when all individuals purchase the risky securities in the same proportion in their portfolio. The conditions under which this can occur are given by the so-called portfolio separation theorem and have been discussed elsewhere.[20] The theorem requires very restrictive conditions on either the utility function and/or the distribution of returns of the set of available securities. In addition, it requires that all individuals have the same expectations about the returns to the firm. It is the latter assumption that we find most objectionable.

---

[16] A third, but unlikely case, arises if there are a large number of firms with perfectly correlated returns (i.e., in the terminology of Modigliani and Miller, they belong to the same risk class). Then if it should happen that there is a firm in the same risk class at the debt-equity ratio our given firm is contemplating, our firm's decisions to choose that debt-equity ratio will not change the set of available securities, even if it involves bankruptcy

[17] See [22] and [23].

[18] That is, borrow funds using the securities as collateral.

[19] Transactions and information costs are clearly important in explaining why conditions (1a) and (1b) are not satisfied.

[20] See Cass and Stiglitz [3].

Finally, transactions costs and information costs associated with bankruptcy also will make the market valuation of the firm depend on the probability of bankruptcy. In the subsequent discussion we shall ignore these costs—in spite of their importance—because we wish to establish that even within the "idealized" world of Modigliani and Miller in which there are no transactions costs, market valuations will in general depend on the debt-equity ratio if there is a finite probability of bankruptcy.

3. The optimal debt-equity ratio with given investment

■ In this section we shall show, by means of a simple example, the dependence of the firm valuation on its financial policy. We postpone until the next section the derivation of the optimal financial policy.

Assume that there are two groups of individuals, denoted by superscripts *a* and *b*.[21] There are two firms[22] in the economy, denoted by subscripts 1 and 2. Group *a* originally owns all the shares in firm 1,[23] the profits of which both groups agree are risky. Similarly, both groups agree that firm 2 is perfectly safe; it has a constant returns-to-scale technology. This fixes the rate of interest on a perfectly safe asset at r^*.[24] We employ the conventional two-period model:[25] investment occurs in the first period, returns occur in the second, and the firm then "dissolves". Our interest here is on the financial decisions, so we assume that the real decisions, i.e., the levels of investment, are already fixed.

We focus our attention on the first firm.[26] It has a return in state θ of $X(\theta)$. The two groups of individuals disagree on the probability of any given state's occurring. We define $F^i(X)$ as the distribution function of X, in group i's judgment. It is convenient to define $\pi^i(X) = 1 - F^i(X)$. Since bankruptcy, in a two-period model, occurs when debt obligations exceed income,[27] i.e., when $X(\theta) \leq (1 + \hat{r}(B)B)$ where $\hat{r}(B)$ is the *nominal* interest rate firm 1 pays if it issues B bonds, $\pi^i[(1 + \hat{r})B]$ is the probability that the firm will not go bankrupt, in i's judgment.

We define

$$\mu^i_{X\geq Y} \equiv \int_Y^\infty X\, dF^i(X). \qquad (1)$$

$\mu^i_{X\geq Y}/\pi^i(Y)$ is the mean of X conditional on X being greater than Y as viewed by group i. In other words, $\mu^i_{X\geq(1+\hat{r})B}/\pi^i((+\hat{r})B)$ is the mean return of the firm when it does not go bankrupt.

---

[21] All of the members of each group are assumed to be identical.

[22] For some purposes it may be preferable to think of these as two industries, since we shall assume that they act competitively.

[23] For the moment, we need not specify how the initial ownership claims are distributed among the members of firm 1.

[24] Provided there is any investment in the second industry at all.

[25] As set forth, for example, in [24]

[26] We accordingly drop the subscript 1 on $X_1(\theta)$, $\hat{r}_1$, etc.

[27] See p. 459 above and [24].

For simplicity, we assume that members of both groups are risk neutral, and that individuals are not allowed to sell shares short. We assume that for all Y, $\mu^a_{X \geq Y} > \mu^b_{X \geq Y}$, i.e., the first group, the original owners of the ith firm, have unambiguously more optimistic expectations about the returns to the firm than do the other group. The return to the shareholders after paying the bondholders is simply

$$\begin{array}{lll} X^i(\theta) - (1 + \hat{r})B & \text{if} & X^i(\theta) > (1 + \hat{r})B, \text{ and} \\ \quad 0 & \text{if} & X^i(\theta) \leq (1 + \hat{r})B, \end{array} \tag{2}$$

where

$$\hat{r} \geq r^*.$$

The mean return is then (in i's judgment)

$$\mu^i_{X \geq (1+\hat{r})B} - (1 + \hat{r})B\pi^i((1 + \hat{r})B) \tag{3}$$

and the value of the equity, E, which is entirely owned by a, is simply

$$E = \frac{\mu^a_{X \geq (1+\hat{r})B} - (1 + \hat{r})B\pi^a((1 + \hat{r})B)}{(1 + r^*)}. \tag{4}$$

We must now determine the function $\hat{r}(B)$; i.e., how does the nominal rate depend on the borrowings of the firm? The second group is risk neutral. Therefore, all that it requires is that (in its judgment) it receives the same average return as on a safe asset (r^*). The return on the risky bond is

$$(1 + \tilde{r}) = \begin{cases} 1 + \hat{r} & \text{if} \quad X \geq (1 + r)B \\ \dfrac{X}{B} & \text{if} \quad X < (1 + r)B. \end{cases} \tag{5}$$

Thus, letting $\mathscr{E}X^b$ be b's expectation of X,

$$\begin{aligned} \mathscr{E}\{(1 + \tilde{r}^b)B\} &= (1 + \hat{r})B\pi^b((1 + \hat{r})B) + \mathscr{E}\{X^b|(X(\theta) < (1 + \hat{r})B)\} \\ &= (1 + \hat{r})B\pi^b((1 + \hat{r})B) + \mathscr{E}X^b - \mu^b_{X \geq (1+\hat{r})B} \qquad (6) \\ &= (1 + r^*)B. \end{aligned}$$

Adding (4) and (6), we obtain

$$V = E + B = \frac{\mu^a_{X \quad (1+\hat{r})B} \geq \mu^b_{X \geq (1+\hat{r})B} - (\pi^a - \pi^b)(1 + \hat{r})B + \mathscr{E}X^b}{1 + r^*}. \tag{7}$$

If the two groups have identical expectations:

$$\pi^a((1 + \hat{r})B) = \pi^b((1 + \hat{r})B) \tag{8a}$$

and

$$\mu^a_{X \geq (1+\hat{r})B} = \mu^b_{X \geq (1+\hat{r})B} \tag{8b}$$

for all B, in which case the value of the firm reduces simply to

$$V = \frac{\mathscr{E}X^b}{1 + r^*} = \frac{\mathscr{E}X^a}{1 + r^*}, \tag{9}$$

so the value is independent of the debt-equity ratio, as we have already argued it would be in this case (since this is just a special form of the mean-variance model).

But note what happens when the two groups differ in their expectations. As long as there is no bankruptcy:

$$\mu^a_{X \geq (1+\hat{r})B} = \mathscr{E}X^a \tag{10a}$$

and

$$\mu^b_{X > (1+\hat{r})b} = \mathscr{E}X^b \tag{10b}$$

so

$$V = \frac{\mathscr{E}X^a}{1 + r^*}. \tag{11}$$

Again, the debt-equity ratio makes no difference. But once

$$B > X^b/1 + r^* \quad \text{for some } \theta,$$

the second group will think the firm may (with a positive probability) go bankrupt. The derivative of the value of the firm with respect to a further increase in the debt is given by

$$\begin{aligned}(1 + r^*)\frac{dV}{dB} = \{\mu^{a\prime}_{X > (1+\hat{r}_1)B} - \mu^{b\prime}_{X > (1+\hat{r}_1)B} \\ - (\pi^{a\prime} - \pi^{b\prime})(1 + \hat{r})B - (\pi^a - \pi^b)\}\frac{d(1 + \hat{r})B}{dB}.\end{aligned} \tag{12}$$

Under our assumptions, the first group is more optimistic than the second, so that even when the second group thinks there is a positive probability of bankruptcy, the first group may not. In particular, defining

$$X^i_{\min} = \min_{\theta} X^i(\theta),$$

if

$$X^a_{\min} > (1 + \hat{r})B > X^b_{\min},$$

the first group thinks that there is no chance of bankruptcy while the second group thinks that there is. Then, at the point where $B(1 + r^*) = X^b_{\min}$;

$$\frac{dV}{dB} = (X^b_{\min} - (1 + \hat{r})B)f^b(X^b_{\min})\frac{d(1 + \hat{r})B}{dB} = 0,$$

where $f(X^b_{\min})$ is the probability of the state where $X = X^b_{\min}$ (in the second group's estimation). But

$$\frac{d^2V}{dB^2} = -f^b(X^b_{\min})\left(\frac{d(1 + \hat{r})B}{dB}\right)^2 < 0.$$

Thus the value of the firm with fixed investment will decrease as it increases its debt beyond that point where the lenders think there is any chance of bankruptcy.

The result that the value of the firm will decrease if the firm has a high debt-equity ratio—sufficiently high that in the view of lenders there is a positive probability of bankruptcy—was derived assuming that there were no real costs involved in bankruptcy. Taking these costs into consideration would reinforce our result. On the other hand, the creation of a "new security" (the risky bonds) might increase the value of the firm if individuals were risk averse (rather than, as we assume here, risk neutral). It is not because the interest rate is an increasing function of the debt-equity ratio that the value of the firm decreases as it issues more bonds, but because there is a divergence in the estimation of the chances of bankruptcy between the lender and the borrower.[28] What limits the amount the borrower borrows is that the amount which the lender must be compensated, using his probabilities, to leave himself indifferent is more than the borrower gains from borrowing, where the borrower is using his own probabilities.

4. Interactions between financial and real decisions

■ In the previous section, we assumed that the level of investment (output) of the risky firm was already fixed, and focused on the financial decision alone. Now that we know that the financial decision may affect the valuation of the firm, it becomes clear that the financial decision may affect the level of investment of the firm. We shall show that maximization of earnings per dollar invested yields both an optimal debt-equity ratio and an equilibrium level of investment in the firm, and that the two decisions cannot in general be separated. In addition, it is noted that the economy may not be productively efficient.

□ **Constant returns to scale.** We assume that there are stochastic constant returns to scale in our first firm: doubling the inputs doubles the output in every state of nature. We further assume that each of the two groups has a fixed amount to invest, I^1; we are concerned with the competitive allocation of this investment. The representative firm in the first industry decides on what percentage of its investment to finance by means of debt. From the analysis of the previous subsection, it should be clear that the nominal rate of interest is simply a function of that percentage, since doubling the number of bonds and doubling the level of investment leaves unchanged the return per bond in any state of nature. A competitive[29] firm believes that it can double its value if it doubles its size, financing a fixed proportion of its growth by bonds and the rest by equity. Hence, because of our assumption of constant returns to scale, competitive equilibrium requires the total value of the firm to be equal to its level of investment. Otherwise the firm would expand or contract.

From our present point of view, however, we are primarily concerned with the choice of a debt-equity ratio. The firm wishes to maximize the expected return per dollar invested by its equity owners, $\mathscr{E}e^1(\theta)$, where

---

[28] In the discussion at Hakone, several participants in the seminar suggested that what was crucial in my analysis was not the bankruptcy assumption but the assumption that the nominal rate of return was an increasing function of the debt-equity ratio. I hope that the above argument convinces the reader that the latter by itself is not sufficient. Moreover, I have not been able to find any convincing argument why the nominal rate of interest should be increasing apart from the considerations discussed here.

[29] The meaning of competitive equilibrium in the presence of uncertainty is not unambiguous, as discussed at length in [22]. As pointed out there, the particular approach taken is, in general, not completely satisfactory, since it does not seem applicable to cases where there are no constant returns to scale (or at least multiplicative uncertainty) and where there is a choice about projects (alternative distributions are feasible).

$$e^i(\theta) \equiv \frac{X^i(\theta) - (1 + \hat{r})B}{E} = \begin{cases} \dfrac{x^i - (1 + \hat{r})\alpha}{1 - \alpha} & \text{if} \quad x^i > (1 + \hat{r})\alpha \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where

$$\alpha = B/I, \tag{14}$$

the proportion of investment financed by bonds, and

$$x^i = X^i/I^i, \tag{15}$$

the return per dollar invested. The probability of not going bankrupt is simply a function of $\hat{r}$ and α. We thus define $\mu^i_{x \geq (1+\hat{r})\alpha}$ and $\pi^i((1 + \hat{r})\alpha)$ just as in Section 3 [equation (1)]: If $G^i(x)$ is the distribution of x in group i's estimation, $\pi^i(1 + \hat{r})\alpha) = 1 - G^i((1 + \hat{r})\alpha)$ is the probability that the firm with debt-equity ratio $\alpha/1 - \alpha$, paying a nominal interest rate of $\hat{r}$, will not go bankrupt, in group i's estimation,

$$\mu^i_{x \geq (1+\hat{r})\alpha} \equiv \int_{(1+\hat{r})\alpha}^{\infty} x dG^i(x).$$

$\mu^i_{x \geq (1+\hat{r})\alpha}/\pi^i((1 + \hat{r})\alpha$ is the mean return to the firm when it does not go bankrupt. The firm wishes to maximize

$$\mathscr{E}e^a(\theta) = \frac{\mu^a_{x \geq (1+\hat{r})\alpha} - (1 + \hat{r})\alpha}{1 - \alpha};$$

that is,

$$\mathscr{E}e^a(\theta) = (1 + \hat{r}) + \frac{d\hat{r}}{d\alpha}\alpha. \tag{16}$$

The value of α, α^*, satisfying (16) is the optimal value of α, and

$$\frac{\alpha^*}{1 - \alpha^*}$$

is the optimal debt-equity ratio.

Equation (16) has a natural interpretation: the average return to equity investment (equals marginal return if equity and debt are expanded proportionately) must equal the marginal cost of borrowing (i.e., of increasing the proportion financed by debt).

If the opportunity alternative to investing in the risky bonds of the first firm for the b group is to buy safe bonds (invest in the safe industry), and if b is risk neutral, then $\hat{r}$ must be such as to make the mean return from holding the first firm's risky bonds equal to $(1 + r^*)$. Denoting the mean return on the risky bonds by $\bar{r}^b$ (in group b's estimation):

$$(1 + \bar{r}^b) = (1 + \hat{r})\pi^b + \frac{\mathscr{E}x^b - \mu^b_{x \geq \alpha(1+\hat{r})}}{\alpha}. \tag{17}$$

Then $\hat{r}$ must be such as to make

$$(1 + \bar{r}^b) = (1 + r^*). \tag{18}$$

Hence

$$\frac{d\hat{r}}{d\alpha} = \frac{\mathscr{E}x^b - \mu^b_{x \geq \alpha(1+\hat{r})}}{\pi^b \alpha^2}. \tag{19}$$

Substituting (15) and (17)–(19) into (16), and letting $y = (1 + \hat{r})\alpha$, we obtain

$$\mathscr{E}x^a = (1 + \hat{r}) + \frac{d\hat{r}}{d\alpha}\alpha = \frac{(1 + r^*)}{\pi^b(y)}. \tag{20}$$

Equation (20) together with (17), which may be rewritten as

$$\alpha(1 + r^*) = y\pi^b + \mathscr{E}x^b - \mu^b_{x \geq y},$$

gives us two equations with two unknowns, (y, α). We can thus solve for the equilibrium debt-investment ratio, α (or debt-equity ratio $\alpha/1 - \alpha$) and the nominal interest on bonds, $1 + \hat{r} = y/\alpha$.

A numerical example may be helpful. Assume b believes that x is uniformly distributed over the interval $[1, 2]$. Then

$$\pi^b = 2 - y, \quad y \geq 1,$$

$$\mathscr{E}x^b - \mu^b_{x \geq y} = \frac{y^2 - 1}{2}, \quad y \geq 1.$$

Let $\mathscr{E}x^a = 1.67$, $1 + r^* = 1.5$. Then

$$\alpha = 0.73$$

$$y = 1.1$$

$$(1 + \hat{r}) = 1.51;$$

that is, equilibrium involves 78 percent of investment financed by debt, a 10-percent probability of bankruptcy, and a nominal interest rate of 51 percent.

Until now, we have assumed that the second group always invests its resources either in the safe asset or in the bonds of the first firm. But they also have the option of buying shares in the first firm. This they will do if the mean return from the shares of the first firm is less than the return on the safe asset or on the bonds of the first firm. The mean return from the shares in the first firm as viewed by the second group $\mathscr{E}e^b$ is given by

$$\mathscr{E}e^b = \frac{\mu^b_{x \geq (1+\hat{r})\alpha} - (1 + \hat{r})\alpha\pi^b}{1 - \alpha}.$$

Subtracting (17), we observe that (after some simplification)[30]

$$1 + \bar{r}^b \gtrless \mathscr{E}e^b \text{ as } 1 + \bar{r}^b \gtrless \mathscr{E}x^b.$$

Thus, whether b prefers investing in the risky bonds of the firm to its shares, depends simply upon whether the bonds pay, in b's judgment, more or less than the total mean return per dollar invested in the firm.

We now are prepared to describe all possible equilibria for this economy. First, it is clear that all the capital of the first group is invested in the equity in firms of the first industry, provided only

---

[30] If the firm issues a sufficiently large number of bonds that even its stockholders believe that there is a finite probability of bankruptcy, then (16) becomes

$$\mathscr{E}e(\theta) = -(\mu^{a'}_{x > (1+\hat{r})\alpha} - 1)\frac{d(1 + \hat{r})\alpha}{d\alpha}$$

with a similar interpretation to that above.

that $\mathscr{E}x^a > 1 + r^*$, the return on the safe asset;[31] if E^a is the value of a's equity in firms in the first industry, and I^a is the amount of resources a has to invest, then

$$E^a = I^a.$$

What the second group does depends on the relationship of $\mathscr{E}x^b$ and $1 + r^*$.

(1) If $\mathscr{E}x^b > 1 + r^*$, then all of the resources of the second group are also invested in the first industry. This will take the form of either bonds or shares. There are two cases. Define α^* as the solution to (16), where $r(\alpha)$ is given by (17) with $1 + \bar{r}^b = \mathscr{E}x^b$ (i.e., investors in the second group must be indifferent between holding bonds and equity):
   (a) If $\alpha^* I^a/1 - \alpha^* < I^b$ then B^b, holdings of the "b" group of risky bonds, are just $B^b = \alpha^*(I^a + I^b)$. The amount invested in equity then will be just $(1 - \alpha^*)(I^a + I^b)$.
   (b) If $\alpha^* I^a/1 - \alpha^* > I^b$ then at a mean return of $\mathscr{E}x^b$, there is an excess demand for borrowing; this drives up the mean return on bonds. The equilibrium then is characterized by the second group holding all of its assets in the form of bonds in the first industry, with a mean return $(1 + \bar{r}^b)$ between the expected return on equity as viewed by the two groups:[32]

$$\mathscr{E}x^b < 1 + \bar{r}^b < \mathscr{E}x^a.$$

(2) On the other hand, if $\mathscr{E}x^b < (1 + r^*)$, then investment in the first industry is given by

$$I^a(1 + \alpha^*),$$

where α^* is again found from (16) where $\hat{r}(\alpha)$ is given by (17) with $\bar{r}^b = r^*$, the safe rate of return. Again, this result is conditional on $\alpha^* I^a/1 - \alpha^* < I^b$. The case where $\alpha^* I^a/1 - \alpha^* > I^b$ follows exactly along the lines sketched above.[33]

The important point to observe here is the interdependence between the real allocation of resources and the financial decisions of the firm. We could not solve for the former until we described the latter. This interdependence is even clearer in the case of diminishing returns to scale.

☐ **Diminishing returns to scale.** We now assume that the returns to the risky firm are described by $X(\theta) = h(I)\varphi(\theta)$, where $h' > 0$, $h'' < 0$, and $\mathscr{E}\varphi = 1$; i.e., there is multiplicative uncertainty but diminishing returns to scale. Bankruptcy occurs whenever $X(\theta) \leq (1 + \hat{r})B$; i.e., whenever

$$\varphi(\theta)\frac{h(I)}{I} \leq (1 + \hat{r})\alpha.$$

---

[31] If this assumption is not satisfied, both a and b invest in the safe industry and the problem is uninteresting.

[32] The exact value is found as follows. From (17). we can solve $\hat{r}$ as a function of $\bar{r}$ and α. Substituting into (16), and letting $\alpha^* = I^b/I^a$, we find that value of $\bar{r}$ which satisfies (16).

[33] That is, if $(\alpha^* I^a/1 - \alpha^*) > I^b$, then $1 + r^* < 1 + \bar{r}^b < \mathscr{E}x^a$.

If $h(I)/I$ is a constant (there are stochastic constant returns to scale), the probability of bankruptcy depends simply on $(1 + \hat{r})\alpha$; but in the case of decreasing returns to scale, at any given debt-equity ratio and nominal interest rate the probability of bankruptcy decreases with the scale of the firm. The consequence of this is that the nominal rate of return on bonds will in turn depend both on the debt-equity ratio and the scale of the firm; the firm will take this into account in making its investment decisions. The result is that the economy will not, in general, be productively efficient; that is, if in the first industry there are several different firms with different production functions,

$$X^j(\theta) = \varphi(\theta)h^j(I), \tag{21}$$

where the j denotes the jth firm in the first industry, efficiency requires that

$$h'^i = h'^j \tag{22}$$

for all i, j in the industry. The market equilibrium will not in general satisfy (22).

To see this, we observe that in equilibrium the return per dollar invested in the equities of all firms in the industry must be the same:

$$\frac{\mathscr{E}(X^j)^a - (1 + \hat{r}^j)B^j}{E^j} = \rho\,;$$

the same is true for all j. Assume that the original shareholders wish the firm to maximize their net worth the first period,

$$\max_{[B^j, I^j]} \frac{\mathscr{E}(X^j)^a - (1 + \hat{r}^j)B^j}{\rho} - (I^j - B^j),$$

where I^j is the value of investment, so $I^j - B^j$ is the new capital raised through issuing new equity. I^j and B^j are chosen so that[34]

$$\frac{\partial(\mathscr{E}X^j)^a}{\partial I^j} = \rho + B^j \frac{d\hat{r}^j}{dI^j}, \tag{23}$$

$$\rho = (1 + \hat{r}^j)\left[1 + \frac{B^j}{1 + \hat{r}^j}\frac{d(1 + \hat{r}^j)}{dB^j}\right]. \tag{24}$$

Equations (23) and (24) must be solved simultaneously: the real investment decision cannot be separated from the financial decision.

Upon solving explicitly for the interest rate $\hat{r}$ which firms have to pay as a function of both the level of investment and the level of debt, (as we did in the previous subsection) and substituting the results into (23) and (24) we can show that (22) will not in general be satisfied. The details are left to the reader.

5. Take-overs

■ **On managers and shareholders.** In the discussion so far, we have assumed that the only nonfinancial decision is the level of investment. We now consider what happens when there is a choice of alternative projects (alternative techniques of production, alternative commodities to be produced, etc.) and what happens if there are disagreements concerning which projects should be undertaken. Our focus is on the consequent possibility of a take-over and the implications this has for the financial decisions of the firm. In order for a firm to make a decision, it must ascertain (1) the probabilities of alternative states; (2) the consequences to the market value of alternative courses of action; and (3) the available set of possible courses of actions.[35]

---

[34] (23) and (24) together imply that the firm is also maximizing its return per dollar invested by the shareholders.

These are all matters in which professionals (in particular, managers) are likely to be more informed than the typical stockholder. As a consequence, it is difficult for the shareholder to judge whether the manager made the "correct" decision. In the event of an unsuccessful project, did the manager assess the probabilities correctly, but have bad luck? Or was the manager inept in judging the distribution of outcomes of the project? Or did the manager choose the best project of the available set, but all the available projects had low returns? Or was the manager pursuing some objective other than that of value maximization (e.g., he was more interested in the size of the firm)?

If we had repeated trials of the same experiment, (the world next year is just as it was last year), where the same opportunities were open to the firm, the manager made the same decision as before, but the outcome, as before, depended on the throw of a die (on the state of nature), then we could discriminate between "bad luck" and "incompetence": on the average, the unlucky manager would yield higher returns to his firm than the inept one. Unfortunately, we do not have repeated trials of the same experiment. Not only are there changes in the set of opportunities available to the firm, but, even if there were not, the outcome of the previous trial will affect our estimate of the distribution of outcomes from the available set of opportunities. Even if the managers of the firm have had a good record in the past, the stockholder must decide, in the instance of a bad outcome, whether this is a matter of bad luck, whether investment opportunities in the industry are "drying up," or whether the manager is no longer "competent" either because circumstances have changed or because the manager has "changed" (because of old age, for example).

Two important conclusions emerge from this analysis. First, unlike conventional neoclassical analysis, in which a firm is nothing more than the factors of capital and land which it owns and the book of blueprints which it confronts and in which the entrepreneur's sole job is to find, at the given factor prices, the correct page of the book of blueprints; here, the entrepreneur plays a crucial role. For instance, the book of blueprints may only be known to him; although factor prices today are given, he must guess what they will be next period. Thus the market value of the firm is not just the market value of its factors (apart from the managers); it depends on the market's estimate of what the managers will do with those factors as well.

Second, individuals may differ not only in their judgment of the probability distribution of the possible outcomes of alternative projects available to a firm but also on their evaluation of the judgment of the managers in making the correct (the "desired") decisions. Indeed, dissatisfaction with the decisions of the firm is not likely to take the form of the shareholders directing the firm to take a particular alternative course of action; rather, it almost invariably consists of replacement of the management. In practice, the instigation for the replacement of a particular management will come not from the shareholders directly, but from an alternative proposed management.[36]

---

[35] If there were a complete set of Arrow-Debreu securities, as soon as one knew the Arrow-Debreu prices, one could easily determine the consequences to the market value and would not have to pay attention to the probabilities of alternative states

☐ **On the determinants of take-overs.** If you own a firm and I observe that you are making the wrong decisions (in my judgment) i.e., you are earning, on the average, a smaller return from your fixed capital than I estimate you could earn (ignoring for the moment any risk aversion), I could offer to buy you out. If this happened everywhere in the economy, all firms would be managed by the individual who (in his judgment) obtained the highest return for that particular firm. However, there are at least three limitations on this process:

(1) Even though I believe I could, on the average, earn a higher return than the present management, my estimates have greater variance than the management's estimates because the management has more "information," and therefore I do not undertake a take-over. This is a formal way of saying that although I think I can do better than the present management on the average, I am not "very sure about it."
(2) Transactions costs. This is a generic name for all the costs associated with a take-over, not the least important of which is the fact that if I fire the manager, he may take with him certain key personnel who have information and whom it would be costly for me to replace.
(3) I may not have the capital to buy a controlling interest in the firm.

Take-overs are of concern to the present shareholders of the firm for at least two reasons:

(1) The possibility of take-overs probably increases the rate of interest which the firm must pay on its bonds. When the original bondholders lend money to a firm, they take cognizance not only of the returns in the different states of nature of the firm under its present management but also of the chance of a take-over bid, a new management which would make an alternative set of decisions. They must take into account all the possible take-over bids, and the resulting dispersion in the possible returns from the firm may be very large indeed. Thus, the return which the original bondholders demand depends not only on the possibility of bankruptcy under the present managements, but also on the possibility of bankruptcy under alternative managements and the probability of these alternative managements' occurring.

---

[36] There are good reasons for this: (1) The decision of whether a particular management had bad judgment is a relative one; it may be "bad," in the sense that each of the shareholders believes he could do better, but of the alternative management teams it may be the best; and (2) The organization of a take-over requires management itself.

(2) Although the majority of shareholders in the firm are better off as a result of the take-over—they would not otherwise have sold their shares—the valuation of the minority may be greater than the market price. They have been, as it were, enjoying some consumer surplus which they will lose if the new management changes its policies. Thus they may be worse off as a result of the take-over.

That firms may wish to avoid take-overs seems clear. The question is, what actions can they take which make it less likely that they will be taken over. Returning to our list of the limitations on take-overs, we observe that the chance of take-overs may be reduced:

(1) By taking actions which make it less likely that an alternative management can argue that it will do a better job; actions which are "risky" *ex ante* always entail the risk of failure, and there is some tendency therefore for firms to take more conservative actions than they might if they were not faced with the possibility of a take-over.[37]

(2) By taking actions which increase the amount that an individual or group must borrow in order to take it over, which decrease the attractiveness of the terms at which they can borrow or the return from the firm once it has been taken over. Three considerations seem to be relevant here.

    (a) For any given total market valuation, the larger is the debt-equity ratio, the more likely is a take-over. To take over control of the firm, the individual needs only to raise enough capital to buy a majority share in the equity. Thus, for a firm with a given market value, a take-over bid is much easier if there is a large debt-equity ratio than if there is a small debt-equity ratio.

    There is one important problem with this argument: if individual borrowing[38] is a reasonably good substitute for firm borrowing, the smaller is the debt-equity ratio of the firm, the more the individual can borrow, to exactly offset the (lack of) borrowing of the firm. There are, however, two reasons why this is not so. Most obvious is the fact that regulations limit the amount an individual can borrow using securities as collateral. Second, bondholders are likely to demand a higher rate of interest from the individual than from the firm, and the rate they demand is likely to be an increasing function of the amount borrowed. The individual may very well have other liabilities, so that the individual may go bankrupt in some states of nature when the firm does not;[39] offsetting this, of course, is the fact that the individual has other assets, so that he may not go bankrupt even when the firm does. Moreover, the lender's estimate of the probability of bankruptcy of the firm depends on who is managing the firm, because now there are real consequences to the control of management. Management may very well insist on a higher nominal rate of return from the individual attempting a take-over than from the original management, for they know less about the decisions he is likely to make with respect to this given firm (even if they have high "hopes" of his doing well, they are likely to have a greater dispersion in their estimate of the outcomes, and hence a higher estimate of the probability of bankruptcy). Of course, it is possible that the individual attempting a take-over has a "good" reputation, in which case he may be able to borrow more cheaply than the firm itself for any given level of borrowing; this will facilitate take-over bids.[40]

---

[37] This cannot be taken to an extreme, for then an alternative management can argue that "good risks" are being passed by.

[38] Since many take-overs are by other firms, the argument is even stronger: we simply substitute direct borrowing of the firm taking over for the "indirect" borrowing of the firm taken over.

[39] Indeed, as a consequence, individuals may be unwilling to borrow extensively, unless the kinds of limited liability arrangements discussed earlier in the previous section exist

(b) If, in the view of those taking over, the nominal interest rate on outstanding loans is lower than would prevail after the take-over (for the level of debt planned by the take-over group), then the high debt-equity ratio increases the attractiveness of the take-over. If, on the other hand, the nominal interest rate is greater than would prevail after the take-over (in the take-over group's estimation) then a high debt-equity ratio (the high interest payments) discourages a take-over.[41]

(c) It is likely that the larger is the firm, the more difficult a take-over will be. Consider two firms identical in every respect (including debt-equity ratio) but that one is twice as large as the other. The group attempting the take-over must mobilize more resources in the latter case than in the former. In particular, the group will have to borrow more. Since the given percentage increase in the requisite debt of the take-over group constitutes an increasing percentage of its total assets (provided that the individual or firm planning the take-over originally had some assets, as it usually has),[42] the larger is the total amount it has to borrow the riskier the borrowing will be in the eyes of the lender and the higher rate it will demand. But, just as in the preceding two cases, the normal presumption may be reversed. If in the view of the lender the firm being taken over has a less risky return than the firm taking over (even taking account of the uncertainty associated with the new management) then the larger is the firm being taken over, the lower will be the interest rate at which the firm taking over can borrow, and the more attractive the investment opportunity. This is typically the case when a small firm attempts a take-over of a larger firm.

---

[40] Indeed, many of the successful take-over bids in the United States by the big conglomerates were based on large borrowings; the market seemed to have a great deal of confidence in these managers. Subsequent events have thrown some doubt on whether this confidence was in fact justified.

[41] One has to distinguish between the interest rates the take-over group must pay in order to succeed in the take-over (which involve the obligations of the company taking over) and the interest rates at which the firm which has been taken over can borrow after the take-over.

[42] The valuation of the assets that is relevant here is the lender's valuation of the worth of the assets at the time that the loan comes due. The net worth of the firm taking over may (from this point of view) be negative.

What emerges from the analysis is that there is no single policy which avoids the possibility of all take-overs; an action which thwarts one take-over bid, from a group with one set of expectations of what it can borrow at, with one set of resources, and with one set of production plans, encourages a take-over bid from another group. Nonetheless, it would seem that there is some presumption that, in general, actions which increase the value of the equity of the firm will make a take-over bid less easily managed. But even here there is an ambiguity. Actions which, in the view of the present management will result in a high market valuation (with a high probability), in the next period, thus making a take-over in the next period less likely, may decrease the market valuation in this period, thus making a take-over in this period more likely.

☐ **An example.** The following example, a slight modification of that presented earlier, shows how the threat of take-over affects the actions of firms.

Assume that there are three groups of individuals. All the members of each group are identical and are risk neutral. Group *a* initially owns industry 1. Industry 1 has two available projects, both of which exhibit stochastic constant returns to scale but differ in the probability distribution of outcomes. In *a*'s estimation, project *a* has a much higher return than the return on a safe asset; in *b*'s estimation it has a zero expected return; and in *c*'s estimation, it has a positive return somewhat higher than r^*. Project *b* has a zero return in *a*'s and *c*'s estimation, and a very high return in *b*'s.

First, we consider the market equilibrium which would result if group *a* and group *c* ignored the possibility of take-over. It would be essentially that described in Section 3. Project *a* would be undertaken. Group *b* would invest in the safe industry. The debt-equity ratio in firm 1 would be

$$\frac{\alpha^*}{1-\alpha^*} = \frac{\gamma I^c}{I^a + (1-\gamma)I^c} \tag{25}$$

(where γ is the proportion of *c*'s investment held in the form of bonds), provided

$$\frac{I^c}{I^a} \geq \frac{\alpha^*}{1-\alpha^*}, \tag{26}$$

as we shall assume. But now consider the possibility of take-over. Group *b* will take over the firm, provided:

(1) His expected return is greater than r^*; since it will be impossible for him to issue bonds, this implies that

$$\frac{\mu^{bb}_{X_1>(1+\hat{r}_1)B_1} - (1+\hat{r}_1)B_1\pi_1^{bb}((1+\hat{r}_1)B_1)}{E_1/2} > 1 + r, \tag{27}$$

where $\mu^{bb}_{X_1>(1+r_1)B_1}$ is *b*'s conditional expectation of the profits of industry 1 *if* project *b* is undertaken. Note that to obtain control only half the shares need be purchased, but once control is obtained, the half remaining in the hands of a shareholder become valueless in a's eyes, and hence can be obtained costlessly if b acted collusively.

(2) It is feasible for them to take over the firm, i.e. their resources collectively are greater than half of the value of the firm:

$$\frac{E_1}{2} < I^b. \tag{28}$$

The value of equity, from (25) is just

$$I^a + (1 - \gamma)I^c,$$

and this may very well be less than I^b, ensuring that a take-over is feasible, and for sufficiently optimistic values of $\mu^{bb}_{X_1 > (1+\hat{r}_1)B}$ it will be desirable.

Although as a result of the take-over b is better off, groups a and c are worse off:

(1) Group c's bonds are now (in its view) essentially worthless.
(2) Although half of the shareholders of group a are indifferent to the new arrangement (since otherwise they would not have sold out their shares) the other half sustain a large capital loss (since in their view the shares are not likely to yield any return and group b may not buy out all the shareholders at the old price of a share). Thus as a group they are worse off.

As a consequence, we have argued that both bondholders and shareholders will take cognizance of the possibility of a takeover. In our example, this will take the following form. Group c will refuse to lend anything if E_1 is less than $2I^b$. This means that there is a sharp discontinuity in the interest rate function. Moreover, unlike the previous case where the nominal interest rate was just a function of the debt-equity ratio, here it is a function of both the debt-equity ratio and the absolute level of equity.[43] Thus the equilibrium debt-equity ratio in our example will be given not by (25) but by

$$\frac{I^a + I^c - 2I^b}{2I^b}. \tag{29}$$

☐ **Further comments on take-overs.** There are two key question areas which have been deliberately avoided in the discussion thus far. The first class of questions concerns the existence of a competitive market equilibrium in economies where take-overs may be important. Indeed, in view of the fact that "control" is vested in 51 percent of the shareholders, the very meaning of "competitive equilibrium" must be called into question. But although it is easy to formulate extended definitions of market equilibria, taking into account the nonmarket aspects of economies with stock markets

---

[43] Recall that in Section 4 we assumed that the firm was acting competitively, in the sense that it assumed that if it doubled its debt and its investment, its market value would double. We expressed some misgivings about this assumption at the time; now that the level of equity as well as the debt-investment ratio affects the interest rate, we must express still stronger misgivings about this assumption.

(the shareholders meetings or the possibility of take-overs by purchase of a 50-percent share), there is nothing to ensure the existence of such equilibria. Consider the following minimal requirements for a market equilibrium:

1. (1) Of all possible plans, the given plan must be preferred by a majority of the original shareholders weighted by shareholdings.
2. (2) The same plan must be preferred by a majority of shareholders after the plan has been announced: now that they have purchased their shares, it must not pay them to hold a new shareholders meeting to adopt an alternative plan.
3. (3) It must not pay any group of individuals, whether original shareholders, shareholders after the plan is announced, or not shareholders at all, to purchase a majority of the shares and adopt an alternative plan: it does not pay any entrepreneur to organize a "holding company" the purpose of which is to buy shares of a given company to take over control, promising to adopt an alternative plan once it has done so.

The requirements may not be satisfied by any plan for a number of reasons:

(1) Except under very stringent conditions, we encounter in both the first and second requirements the usual voting problems: there may exist no plan which is preferred against all other plans by a majority (cyclical majorities). But even if there were only two groups of individuals in the economy, we would encounter difficulties.

(2) Consider, for instance, a situation where there are two possible production plans for the firm, one of which in one group's estimation has a very high return, but in the other group's estimation has a very low return, and conversely for the other plan. Since only a majority of the shareholders need to be "bought off," it is clear that if the firm is controlled by the first group, the second group can "buy off" a majority of the first group, and if the second group controls the firm, the first group can "buy off" a majority of the second group. Thus the third requirement is not satisfied. Only by acting collusively can such take-overs be stopped.

(3) Indeed, not even a formal take-over is required for these difficulties to be encountered. In the above example, if originally the firm is owned by the first group, as a result of the high expected return of the project (accompanied by a relatively low variance) the individual will have a high level of expected utility. Accordingly, the individual may allocate a larger proportion of his assets to the safe firm than he would have allocated if the return were low (at the same variance). It is clearly possible then that when the first group originally owns a majority of its shares and adopts its preferred project, with a high expected return, it ends up a minority shareholder, and conversely when the second group originally owns a majority of the shares. Thus the first and the second requirements may not be satisfied simultaneously.

(4) Still a fourth difficulty is encountered when there is more

than one factor or commodity. To establish existence of equilibrium, the excess demand curves for the different commodities and factors need to be continuous (or upper-semicontinuous). Production today depends on expected prices in the different states of nature in the next period, which in turn depend on production plans tomorrow. Normally, we would expect slight changes in prices today to have slight effects on production plans and expected prices tomorrow, and hence to have a slight effect on production plans undertaken today—and accordingly only a slight effect on excess demands today. But if as a result of a slight change in the price, the firm goes from having a profit equal to its obligations to bondholders to having less than enough to pay its bondholders (it goes bankrupt), then the decisions taken by the new management group —whose expectations about returns in the future for alternative projects may be quite different from those of the original management—are likely to be discretely different from those taken at the previous price. But this means that there is no basis for arguing that the excess demand curves in this period will be continuous.

We have only been able to hint here at the difficulties involved simply in establishing the existence of equilibrium for a market economy exhibiting a few of the features which we seem to observe in modern capitalist countries.[44] I have explored some nonmarket concepts of equilibrium, in particular, the core, but these seem equally unsatisfactory. Indeed, we can construct examples where the core is empty.

The second class of questions has to do with information, the determination of expectations of returns, and the value of a firm. The return to an individual from holding a share in a firm is not only the dividends but also the capital gains. Thus the price today depends on the expected price tomorrow. The expected price tomorrow depends not so much on what we think the firm's present discounted value of returns will be tomorrow, but what the market assesses as its value. That this may give rise to speculative booms is clear and has been discussed elsewhere. It has been shown that (in a deterministic model) "every speculative boom must come to an end:" in finite time the price must become infinite, and somebody winds up holding the shares when the boom busts. One might argue that the market would "learn" to avoid such behavior; but without futures markets, there seems to be no mechanism for avoiding these difficulties. What I want to point out here are the further difficulties arising in the presence of uncertainty. For when the share no longer yields the high returns expected, the individual must decide whether (1) he was unlucky, in the sense that he recognized that in some states of nature the price would be low, and ascribed a low probability to the occurrence of those states, but nonetheless they sometimes occur; (2) further information has become available affecting other individuals' expectations about returns in the future and hence the price of the share; or (3) the original price was "too high." There are no sure grounds for differentiating among these.

---

[44] And, perhaps equally important, not making use of the Arrow-Debreu securities which do not in fact exist

The fact that individuals must rely in part upon other's estimates of the value of the firm has two important implications for take-overs.

First, the fact that an individual or group of individuals estimates that the result from controlling this company is greater than its present price is information to the stockholders; not unlikely, as a consequence, they will revise upwards their estimate of the returns to the firm. Indeed, if the firm or individual attempting the take-over offers only slightly more than the present market price, the stockholders may revise their estimate of the market value as a result of his offer so that it now exceeds his offer. That is at least part of the reason why take-over bids are almost all substantially above the market price of the firm at the time they are made.

Second, in borrowing, the lender may have little information about the probability of bankruptcy. If he observes like-minded people insisting on the more than safe return, he may use this as evidence that others think that there is a large probability of bankruptcy. Thus he may revise his estimate of the probability of bankruptcy, insisting on an even larger nominal return. The effect of all of this is to make the supply curve of loan capital very sharply upward sloping beyond some point, and make it more difficult for individuals and firms to acquire the funds required for a take-over by borrowing.

6. Concluding comments

■ We have shown in this paper that the possibility of bankruptcy has very strong implications for firm behavior: it may not be possible to separate out financial and real decisions and there may indeed be an optimal debt-equity ratio. Moreover, we have demonstrated the further consequence that the economy may not be productively efficient.

Some economists have argued that because bankruptcies are relatively infrequent, they cannot be important; that a model which assumes they cannot occur is more "realistic" than one, as presented here, in which they play a central role. We would argue that even if bankruptcies are relatively infrequent, the possibility of bankruptcy may have an important effect on firm behavior; for instance, they may choose a debt-equity ratio which avoids having a very high probability of bankruptcy. Moreover, one ought to include some take-overs and mergers in the same category as bankruptcy; rather than go through the high costs associated with actual bankruptcy, a firm which is approaching insolvency is likely to prefer these methods of disappearance. These have been quite frequent both in the United States and the United Kingdom in recent years.

We would argue that in the models we have investigated here, not only are the assumptions more realistic than those of the conventional models but the predictions of the models seem more in accord with how firms behave. In any case, further exploration of models with bankruptcy and divergent expectations seems called for.

References

1. Arrow, K. J. "Aspects of a Theory of Risk-Bearing." Yrjo Jahnsson Lectures, Helsinki, 1965.
2. ———. "The Role of Securities in the Optimal Allocation of Risk-Bearing" *Review of Economic Studies,* Vol. 31, No. 2 (April 1964), pp. 91–96.
3. Cass, D. and Stiglitz, J. E. "The Structure of Preferences and Returns, and Separability in Portfolio Allocation: A Contribution to the Pure Theory of Mutual Funds." *Journal of Economic Theory,* Vol. 2, No. 2 (June 1970), pp. 122–60.
4. Debreu, G. *Theory of Value. An Axiomatic Analysis of Economic Equilibrium.* New York: John Wiley & Sons, Inc., 1959.
5. Diamond, P. A. "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty." *The American Economic Review,* Vol. 57, No. 4 (September 1967), pp. 759–76.
6. Hahn, F. H. "Equilibrium Dynamics with Heterogeneous Capital Goods." *Quarterly Journal of Economics,* Vol. 80, No. 321 (November 1966), pp. 633–46
7. ———. "Equilibrium with Transactions Costs." *Econometrica,* Vol. 39, No. 2 (May 1971), pp. 417–40.
8. Hirshleifer, J. "Investment Decision under Uncertainty: Applications of the State-Preference Approach." *Quarterly Journal of Economics,* Vol. 80, No. 2 (May 1966), pp. 252–77.
9. ———. "The Private and Social Value of Information and the Reward to Inventive Activity." *The American Economic Review,* Vol. 61, No 4 (September 1971), pp. 561–74.
10. Jensen, M. and Long, J. B , Jr. "Corporate Investment under Uncertainty and Pareto Optimality in the Capital Markets." *The Bell Journal of Economics and Management Science,* Vol 3, No. 1 (Spring 1972), pp. 151–74.
11. Knight, F. H. *Risk, Uncertainty and Profit.* Clifton, N. J.: Augustus M. Kelly, 1921.
12. Lintner, J. "The Aggregation of Investor's Diverse Judgements and Preferences in Purely Competitive Security Markets." *Journal of Financial and Quantitative Analysis,* Vol. 4, No. 4 (December 1969), pp. 347–400.
13. ———. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *The Review of Economics and Statistics,* Vol. 47, No. 1 (February 1965), pp. 13–37.
14. Modigliani, F. and Miller, M. H "The Cost of Capital, Corporation Finance and the Theory of Investment." *The American Economic Review,* Vol 48, No. 3 (June 1958), pp. 261–97.
15. Mossin, J. "Equilibrium in a Capital Asset Market." *Econometrica,* Vol. 34, No 4 (October 1966), pp. 768–83.
16. Radner, R. "Competitive Equilibrium under Uncertainty." *Econometrica,* Vol. 36, No 1 (January 1968), pp. 31–58.
17. Rothschild, M. and Stiglitz, J. E. "Increasing Risk: A Definition and Its Economic Consequences." *Journal of Economic Theory,* Vol 2, No. 3 (September 1970), pp. 225–43, and Vol. 3, No. 1 (March 1971), pp. 66–84.
18. Sharpe, W. F. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk" *Journal of Finance,* Vol. 19, No. 4 (September 1964), pp. 425–42
19. Shell, K. and Stiglitz, J. E. "The Allocation of Investment in a Dynamic Economy." *Quarterly Journal of Economics,* Vol. 81, No 325 (November 1967), pp. 592–609.
20. Stiglitz, J. E. "The Degree of Product Differentiation in Monopolistic Competition: An Example." Mimeographed. The Cowles Foundation, Yale University, 1972.
21. ———. "On the Irrelevance of Corporate Financial Policy" Cowles Foundation Discussion Paper No. 339, Yale University, April 1972.
22. ———. "On the Optimality of the Stock Market Allocation of Investment" *Quarterly Journal of Economics,* Vol. 86, No. 342 (February 1972). pp. 25–60.
23. ———. "Perfect and Imperfect Capital Markets." Paper presented at the meeting of the Econometric Society, New Orleans, 1971.
24. ———."A Re-Examination of the Modigliani-Miller Theorem." *The American Economic Review,* Vol. 59, No. 5 (December 1969), pp. 78–93.
25. ———. "Taxation, Corporate Financial Policy, and the Cost of Capital" *Journal of Public Economics,* forthcoming.

26. ______. "Value Maximization and Alternate Objectives of the Firm." Mimeographed. The Cowles Foundation, Yale University, 1972.
27. Stigum, B. "Competitive Equilibrium under Uncertainty." *Quarterly Journal of Economics,* Vol. 83, No. 333 (November 1969), pp 533–61
28. Tobin, J. "Liquidity Preference as Behavior toward Risk." *Review of Economic Studies,* Vol. 25, No. 1 (February 1958), pp. 65–86.

FCC regulation, monopsony, and network television program costs

Robert W. Crandall

Associate Professor of Economics
Massachusetts Institute of Technology

In this paper, a television program cost function is derived from a simple model of network rivalry based upon the Cournot postulate. This cost function is then fitted to data from a cross section of new network entertainment series for 1960–1965. The resulting estimates are quite unfavorable to two different aspects of FCC policy towards television broadcasting.

First, the Commission's implicit attempt to create monopoly profits for networks and their affiliates in entertainment programming for the purpose of subsidizing nonremunerative "public interest" programming is shown to be extremely inefficient. As much as two-thirds of entertainment program costs are simply rents paid to star performers and other inputs to the program production process—rents which derive from the Commission's allocations policy.

Second, the Commission's recent decision to bar network investment in non-network rights to programs first exhibited on their interconnected stations is demonstrated to be unnecessary. The Commission has enunciated this policy in its recent Prime Time Rule in order to protect program suppliers from monopsonistic exploitation by the networks. But the cost-function estimates show that the networks have paid their suppliers the economic value of the non-network rights obtained in the initial bargaining process. The only effect of prohibiting network investment in these ancillary rights is to limit the ability of suppliers to sell risky assets.

1. Introduction

■ The Federal Communications Commission's recent admission that it is unable to regulate interstate common carriers of communications services may strike the casual observer as shocking, but it is no more than a candid admission of the problems facing this commission and others as they attempt to render precise judgments about market relationships which they cannot measure. Whether it is the Federal Power Commission regulating the supply

---

Robert W. Crandall received the B.A. degree in economics from the University of Cincinnati (1962) and the M.S. and Ph.D degrees in economics from Northwestern University (1965 and 1968, respectively). Dr. Crandall is currently Associate Professor of Economics at M.I.T. and serves as a consultant to the Urban Institute in Washington, D C. In addition to his continued research in cable television broadcasting, he is developing an econometric model of labor markets.

Research support for the empirical research reported in this paper was provided by a Post-Doctoral Fellowship in Public Utility Economics granted by the American Telephone and Telegraph Company. Additional support was provided by the Department of Economics at M I.T. The data utilized in this paper were made available by the three network companies, but the conclusions reported herein are those of the author and do not necessarily reflect the opinions of the networks. Generous advice on matters of econometric estimation was provided by Professor Franklin M. Fisher, but responsibility for all errors remains with the author.

of natural gas,[1] the Interstate Commerce Commission regulating surface transportation,[2] or the FCC regulating television broadcasting, a common problem looms. Without precise empirical knowledge of price-cost relationships, these regulatory bodies must sit as puzzled arbiters of conflicting claims made by various private and public interest groups.

One might think that the FCC requires little precise knowledge about television markets, since it does not regulate the networks' or stations' profits directly by limiting the prices charged to advertisers or by prescribing a minimum output level. Rather, the Commission focuses upon performance in "the public interest," which is variously defined as providing the maximum diversity or quality of television programming. But price-cost relationships are no less important in this pursuit than in the attempt to regulate common carriers, for the Commission has followed a policy of subsidizing nonremunerative "public-service" programming by protecting local stations and networks from new sources of competition. Severe restrictions upon "pay" television,[3] limitation on the growth of cable television activity,[4] and even quantitative restrictions upon network broadcasting during prime time are enacted principally to protect local television stations from an erosion in their profits—profits which the Commission wishes to see devoted in part to public-service programming. But most restrictions upon competition at the local level result in the limitation of the number of national networks as well, for it is the paucity of local affiliates which has prevented entry into network broadcasting for two decades.

Unfortunately, what has not occurred to the Commission is the possibility that its attempt at cross-subsidization has largely failed and that it has simply conferred large rents upon performers and other persons involved in producing network entertainment programming. Nonprice rivalry among networks and local stations is likely to result in dissipation of the monopoly rents which the Commission wishes to allocate to public-service programming, with the result that viewers are forced to contribute to the high salaries of television celebrities by sacrificing popular programming choices.

The most recent indication of the Commission's failure to grasp the nature of competition and nonprice rivalry in network broadcasting and program supply is to be found in its 1970 Prime Time Rule.[5] This decision limits network broadcasting in "prime time" (7:00 P.M. to 11:00 P.M., Eastern time) to three hours per day in the nation's largest 50 markets and prohibits network investment in subsequent non-network distribution of programs first exhibited on national television. The latter prohibition is designed to protect program suppliers from network monopsony power, but the FCC has no evidence that such power actually exists.[6] Even more troubling is the Commission's discussion of program "costs" in defending its rule-making, for there appears to be little recognition that these expenditures are not exogenously determined social opportunity costs.

---

[1] See MacAvoy [4] for a rather glum appraisal of the FPC's performance.

[2] Among the almost universally critical studies of the ICC, the most recent treatise is that of Friedlaender [2].

[3] The Commission's attitude toward pay television is chronicled in the numerous decisions rendered under F.C.C Docket 11279

[4] See Final Cable Television Decision [6].

[5] See *Report and Order* [9].

Clearly, if the FCC is to pursue its policy of cross-subsidization, it could benefit considerably from careful research into the nature of network and local station rivalry and the determinants of programming expenditures. This paper represents a modest start in that direction, for in it we seek to provide an econometric estimate of a network program cost function by utilizing actual cost data from the three national commercial networks for 1960–1965.[7] Included in the program cost function are variables which measure the extent of network investment in non-network exhibition rights of programs first shown on the networks. The regression coefficients for these variables may be compared with financial returns from program distribution in order to determine the extent to which networks exercise monopsony power over their suppliers. In addition, we are able to derive from these regressions estimates of the magnitude of the economic rents which are siphoned off by performers and other inputs to the production of programs. Finally, we utilize the program cost regressions to draw some inferences about the price elasticity of advertising demand facing the networks.

2. Network markets

■ **The product market.** A network is in large part a broker of its affiliates' time. It contracts with these affiliates to sell time to national advertisers who may conceivably bring their own programming to the network for exhibition. In practice, the networks have also become program brokers, selling not only time but viewers.[8] Each network schedules its programs and offers commercial announcements at regular intervals during these program hours.

Since there are three or fewer competitive commercial stations in most local markets, there are only three national brokers. At present, a fourth broker would be able to offer advertisers fewer viewers at each hour for programs similar to those broadcast by its three rivals. Given the current profit margins (per unit sales), it is doubtful that a fourth network could survive with this disadvantage.

Despite the entry barriers provided by regulation of the spectrum and fewness of rival firms, network companies do not necessarily enjoy monopoly power in the national advertising market. Given the numerous substitute media—newspapers, radio, magazines, and local television advertising time—it is distinctly possible that variations in a network's output of viewer hours have no effect upon the price of viewer hours in the market. But even with a network television market demand curve, which is infinitely elastic at a per viewer exposure price set by competition from other media, the revenues per period for each network are a function of the number of rivals in network broadcasting. Unless the total number of viewers per hour increases in proportion to the number of rival networks—an unlikely possibility—any increase in the number of networks must decrease revenues per hour and, therefore, total revenues per year.[9]

---

[6] The informal, oral evidence heard by the Commission's staff is reproduced in two Interim Reports on television network program procurement [7] The second of these reports will be referred to hereafter as the Second Interim Report.

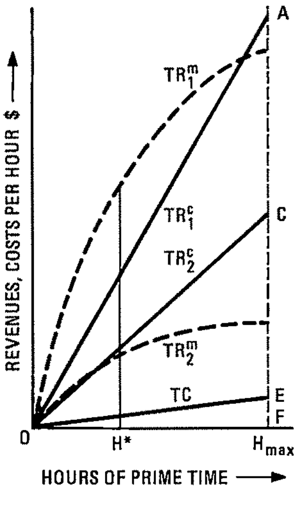
[7] The years 1960–1965 were chosen because data for this period were most accessible when the author began to collect data for this study. Moreover, older data would be less valuable simply because network advertising sales were undertaken in a different fashion before the 1960s.

[8] For a detailed description of this development, see Crandall [1].

If we assume that the number of viewers per hour of prime viewing time is constant, that they divide equally among the networks' offerings, and that the number of prime-time hours per year is fixed at $H_{\max}$, the network's revenue function may be depicted quite simply as in Figure 1. If there is but one national network facing a competitive advertising market, its total revenue function, TR_1^c, will be linear with a slope AF/OF which is equal to revenues per hour; but the appropriate measure of a network's output is the number of viewer exposures to commercial messages of a given length. If the number of these messages is fixed per hour of broadcasting, output can be expressed in viewer hours of programming. The price of these viewer hours in a competitive advertising market is determined by the price of equivalent advertising messages in alternative media. Therefore, in this competitive setting, the network's revenues per hour are exogenously determined by the number of viewers watching television during each time period and the cost of advertising in the other media. The entry of a second network with access to equivalent affiliates and similar program fare will simply reduce each network's total revenue by one-half and shift its revenue function down to TR_2^c, where revenues per hour are equal to CF/OF. If program costs are invariant against changes in average audience size, each network incurs programming costs combined with brokerage, interconnection, and affiliation costs, which sum to a total cost per hour of TC. In this hypothetical world, entry could continue until the TR schedule of each network fell to a slope of EF/OF.

FIGURE 1

GROSS ADVERTISING REVENUES FOR A SINGLE NETWORK DURING PRIME TIME



Under these restrictive assumptions about viewer and cost behavior, network monopoly power in advertising markets simply alters each network's total revenue function. For instance, a single network's schedule becomes TR_1^m under the assumption that it faces the same price per viewer exposure as a competitive firm at an output corresponding to H^* hours of programming. Entry of a second network shifts the total revenue schedule down to TR_2^m, under the assumption that each firm chooses to broadcast the full $H_{\max}$ hours per year. The limit to entry occurs when the appropriate TR schedule is just tangent to the TC schedule. Obviously, this entry limit is reached sooner if networks enjoy monopoly in their advertising market than if they are perfect competitors in this market.

---

[9] At present, over 60 percent of the nation's homes have their receivers switched on during prime time. Clearly, this audience could not swell appreciably with the development of new viewing options.

We see then that the effect of "monopoly" power is somewhat different in network television from that in most normal markets. For instance, if entry is so effectively impeded by allocations policy as to allow only, say, two national networks, profits may be earned by each network even though its demand curve has infinite elasticity. Given the assumption of fixed audience sizes during prime time, monopoly power could only be exercised by limiting the number of hours broadcast. Thus, in this model, the existence of network monopoly power in advertising markets is quite unimportant except as it provides a limit to the number of networks which could compete successfully. Once we relax the assumption of fixed program costs, however, this conclusion is modified but not completely negated.

☐ **The input markets.** The preceding discussion abstracts rather heroically from reality. A television network does not face a fixed audience, but rather it may be assumed that its audience will vary with its choice of programming. Indeed, one might argue that the most important management decision in network broadcasting is the choice of programs and the scheduling of these programs against rivals' offerings. In addition, the network's affiliation costs —the prices it must pay for access to its affiliates—cannot be assumed to be given exogenously. We may dispose of the affiliation costs first, since the principal focus of this paper is upon program choice and program supply.

An important source of network rents is the limitation upon local station spectrum allocations. The FCC has allocated the spectrum to television in such a manner that no more than three competitive commercial stations may operate in a great number of markets. In those markets in which a fourth station may obtain a license, it is frequently limited to the UHF band. But a number of markets have four or more competitive local stations, and this inconsistent allocation provides the best of all possible worlds for the three existing networks. Without a fourth station in virtually every large market, a fourth network cannot hope to enter profitably, but the presence of four stations in many of the largest markets allows the networks to appropriate a large share of the rents from network broadcasting. With more stations than potential affiliations, cartel-like agreements among stations are not likely to develop in bargaining with networks, simply because there is no simple method for the winners to compensate the losers in the affiliation competition. Were there only three stations in every market, the local stations would certainly appropriate a larger share of the net revenues from the sale of network advertising.[10]

More important for this paper's analysis is the nature of program supply and the networks' position as purchasers of programming, especially entertainment programming. With the trend toward direct network negotiation with program suppliers came the possibility that networks might obtain monopsony power in program supply. In earlier years, a number of advertisers licensed their own programs and brought them to the networks in return for payment for an entire time period. In the past decade, however, this procedure has practically vanished for prime-time entertainment programming. For all intents and purposes, it is the networks who choose the programs in which advertisers intersperse their messages; but this is hardly sufficient proof that they possess monopsony power over the price of programs.

---

[10] For an econometric estimate of the effect of local station competition for affiliation agreements, see Robbins [5]. Robbins' study reveals that the networks pay significantly more per home reached by affiliates in markets with three or fewer stations than they do in markets with four or more stations. Of course, a limit to the difference in compensation rates is provided by the threat of entry of a fourth network.

Entertainment programs are simply filmed or taped products which are exhibited on television rather than in alternative media. They require virtually the same inputs as motion pictures or other films and may be produced by any of scores of production companies throughout the world. Thus, television networks must bid resources away from alternative media in order to obtain their required program inputs. Given the increasing number of buyers of filmed products, it is quite clear that the networks' monopsony power must be limited.

Even if the three networks were the only buyers of filmed entertainment output, it would be necessary for program supply to be increasing cost for them to exert monopsony power. And if this monopsony power were exerted, it would become evident in the networks' failure to offer programming during some periods of considerable viewing frequency. Clearly, this last condition is not satisfied and has not been in the past few years. Before FCC restrictions on the length of the network program day, there were only occasional half hours during prime time which were not filled, and most of these periods were on A.B.C., which continues to have a comparative disadvantage in affiliations. Certainly, A.B.C. was not exerting monopsony power while N.B.C. and C.B.S. struggled along as competitors for programs!

But a more fundamental objection to the standard monopsony question obtains: it is impossible to draw a supply curve for homogeneous programs because programs are far from homogeneous. Networks are not buying standardized items from the supplying companies; rather, they are selecting the appropriate program input mix and contracting with suppliers to supervise their combination into programs. Given the number of such "packagers" of programs, the networks need only promise them a return sufficient to cover their costs, including the costs of invested capital. They may pay no less; they need pay no more.

The quantity and value of resources employed by the networks in their entertainment programming depends upon the degree of competition for audience and the price per viewer which advertisers are willing to spend. The greater are program revenues, the greater will be network expenditures upon programming. Networks will choose to employ more reproducible resources per program; and nonreproducible inputs, such as star performers, directors, and writers, will obtain higher rents when program revenues are high. While increasing the number of networks may increase the percentage of revenues diverted to programming expenses, the expenditures per program must certainly fall and with them the rents to talent inputs.

But even if program costs vary directly with revenues, it is still possible that the condition of blockaded entry in network broadcasting finds each network with a total revenue function which is above the total cost function in Figure 1. All we have argued thus far is that the TC schedule will shift downward with the TR schedule as the number of networks increases, but TC will fall less rapidly than TR if nonprice rivalry increases with the number of competing networks. This is neither new nor profound, but it serves to emphasize two points: (1) neither monopoly power in advertising markets nor monopsony power in programming markets is required to yield the networks a substantial scarcity rent; and (2) the attempt by the FCC to create large excess profits in network entertainment programs in order to finance local and network "public service" programs may be frustrated by the nonprice competition which results. The latter point is especially important if program rivalry results in large increases in program costs which are not valued highly by viewers. To the extent that this rivalry simply bids up performers' rents, the profits are clearly wasted. If the rivalry takes the form of attracting scarce resources from other markets, an additional social cost results which must be weighed against the increase in viewer satisfaction. But, in any case, unless the FCC acts to curb this type of nonprice rivalry, its implicit attempt to create monopoly profits for the purpose of subsidizing other services is surely weakened.

☐ **Residual rights.** Few network programs are fully exploited during their network run. In practice, each network licenses the right to exhibit a program for one or more exhibitions, and if this program is recorded on videotape or film it may be shown subsequently in other markets. This further exploitation of the program is referred to as "syndication," and it assumes two forms. First, the program may be licensed to foreign markets almost simultaneously with its United States network exhibition. Such foreign syndication accounted for $57 million in revenues in 1967.[11] In addition, after the network program series has concluded its run, the program may be licensed to United States television stations for additional local exhibitions. These domestic "off network" syndication sales accounted for $43.1 million in total revenues in 1967.[12]

The success of a program in syndication sales to local stations and to foreign markets is a direct function of its success during the initial network run. If it has a long run on the network, it will probably be quite successful for two reasons: (1) it has a demonstrated audience appeal; and (2) there are a number of episodes available for distribution. This latter factor is extremely important, because it is uneconomical to produce additional episodes once a series has ended its network run. In addition, a television station is much less interested in procuring a series for which there are only 20 or 30 episodes than it is in one which has hundreds of episodes.

In recent years, the three domestic networks have purchased approximately 90 percent of their prime-time entertainment series from independent "packagers" of programs.[13] As networks turned to direct licensing of programs from independent suppliers, they obtained in the initial bargaining process rights to a share of the programs' future syndication revenues as well as interests in the revenues which might be generated by any exploitation of copyrighted matter from these programs in the sale of merchandise. Our primary interest lies only in syndication revenues, however, since merchandising revenues are less than 5 percent of the syndication yield.

---

[11] See Little [3], p. 109, hereafter referred to as the Little Report.

[12] *Ibid.*

It is crucial to understand that network syndication interests have always been obtained during the initial bargaining process between the program supplier and the network. Before the program series reaches the network, an agreement is reached between the network company and the supplier as to the price which the network will pay for each episode in the series, the price to be paid for repeat episodes on the network, and the shares of syndication rights and profit shares to be retained by the network, as well as a myriad of other matters such as allowances for unexpected cost increases in producing the programs, escalator clauses for subsequent seasons, the division of merchandising interests, penalties to be paid in case of network cancellation of the series, and technical details concerning delivery of episodes to the network. In most cases, the network obtains the rights to the network exhibition of the program series for five years.

The rights to distribute program episodes to domestic television stations after the initial network run and to foreign stations during and after network exhibition are referred to as domestic and foreign syndication rights. The right to share in the net revenues from syndication—equal to total revenues less distribution costs, entertainment guild royalties, and other costs—is referred to as a profit share, and this share is usually identical for domestic and foreign revenues. Prior to the 1970 Prime Time Rule, the initial contract between network and packager not only allocated the above interests, but it also specified the precise distribution fee to be charged by the distributor. These fees ranged from 10 percent to 45 percent, the larger fees being exacted for the smaller sales. In addition, networks often agreed to allow packagers to recoup excess production costs from the networks' profit shares.

While the syndication market is not highly concentrated, one may not assume that syndication rights and profit shares are not valuable. The fee stipulated in the original network-packager contract may be sufficiently large to compensate the distributor more than amply for his costs of distribution. Moreover, the actual costs incurred in distribution are very small since the episodes have already been produced. Only shipping, print, and other miscellaneous costs such as royalty payments to members of entertainment guilds are deducted from the revenues which remain after imposition of the distributor's fee. Thus, syndication revenues are an important contributor to net profits even if they are small in comparison with the original network payments for the programs.

According to available data,[14] networks obtained the syndication rights in one-fourth to slightly more than one-third of their prime-time programs from 1957 through 1968. This proportion reached its maximum in 1964—35 percent for domestic distribution rights and 35.7 percent for foreign rights—but it declined steadily to 25.9 percent for both in 1968.[15] The networks' share of actual syndication sales was estimated at 22 percent of the domestic market and 29 percent of the foreign market for 1967, the most recent year for which data are available.[16] Although neither datum is likely to indicate substantial market power, the FCC's Prime Time Rule has imposed a total prohibition upon network distribution in domestic syndication markets. As we shall see, this decision was motivated as much by the FCC's fear of network monopsony power as by its concern about undue concentration in syndication.

---

[13] *Ibid.*, p. 5.

[14] *Ibid.*, p 46. The FCC Office of Network Study also collects these data.

3. The FCC rule

■ In 1965, the FCC issued a *Notice of Proposed Rule-Making*[17] which sought to prohibit: (1) all domestic syndication (distribution) by network companies; (2) network ownership of syndication profit shares in programs produced wholly or in part by others; and (3) foreign syndication by the networks of programs produced wholly or in part by others. In 1970, these provisions formally adopted by the Commission in its Prime Time Rule, whose principal thrust is to limit the number of hours of prime-time network broadcasting in the nation's largest 50 markets.[18]

Why should the FCC act to prohibit network purchase of these rights from packagers? Even if it fears monopoly control over domestic syndication, is it necessary to prohibit networks from buying even a minority interest in the profits from syndication? Prohibition of network distribution might be warranted, but it is surprising that the Commission has chosen to limit suppliers' right to sell profit shares in their productions.

A reading of both the Proposed Rule and the Prime Time Rule reveals that the Commission believes that it is not restricting suppliers' prerogatives but rather protecting these packagers from the monopsony power of the networks. By prohibiting the networks from bargaining for syndication profit shares and distribution rights, the Commission is reserving these lucrative privileges for the suppliers. The Commission reasons that the networks have been utilizing their monopsony power to bid these profits away from suppliers at little or no compensation. As a result, program producers were courting bankruptcy, and causing a threat to the continuance of many sources of program supply. The Commission's implicit theory of network-producer bargaining may be inferred from various statements appearing in the *Notice of Proposed Rule-Making*. For instance, it is claimed that:[19]

> . . network corporations—usually as a *quid pro quo* for initial financing but sometimes as compensation for assumption of the risk of sale to advertisers— . . . seek and frequently obtain . the right to distribute the programs or series in domestic syndication and foreign markets, the right to share . . in the profits from domestic and foreign syndication sales . .

---

[15] *Ibid*

[16] (Percentages of off-network sales only.) *Ibid.*, pp. 108–9.

[17] [8].

[18] See [9]

[19] [8], p 8.

But these rights were being ceded unwillingly by producers, since in earlier years:[20]

> [t]hese rights usually were retained by independent producers and constituted valuable commercial assets which contributed to their economic stability and viability The importance of the retention of these rights to the financial stability of independent producers is supported by testimony of producers that in many, if not most, instances they do not recover their initial production costs from the network run . . . but must look to syndication and foreign sales to 'make them whole' and to show a profit.

There can be little doubt that the Commission believed that its Proposed Rule would save independent producers from network exploitation, for the Commission stated:[21]

> . . . [I]t is our hope that the proposed rule would reduce the possibility that independent producers may be *forced* to give up rights in their programs in order to obtain access to network time. . .

What did not occur to the Commission is the possibility that independent producers "gave up" rights in their programs because networks offered to buy them. The Prime Time Rule allows networks to sell the rights in the programs which they produce for themselves, but ignores the possibility that transfers of rights in the other direction are equally compensated. This is not an unusual posture for a regulatory agency which finds itself enmeshed in transactions between very large and very small firms. It is a common regulatory notion that large firms exploit the smaller firms simply because they are larger, but more often than not this theory does not stand up to careful scrutiny.

The alternative theory which is tested below is very simple. It is merely that networks purchase ancillary rights in programs exhibited originally on their networks at prices which reflect *ex ante* expectations of their discounted present value. If a network wishes to invest in either syndication rights or profit shares, it must compensate the supplying company through correspondingly higher initial prices for episodes of the program which are shown on the network. Therefore, program prices to the networks should vary directly with the magnitude of their syndication interests in each.

4. A model of network programming

■ It is impossible to posit program-supply function without also addressing the problem of network rivalry and network-supplier negotiations for residual rights. The value of a program is directly related to its ability to attract audiences, but actual audience is inversely related to the strength of competing programs at the same hour. Moreover, the number of viewers attracted to a program will vary systematically throughout the day and even during prime time. Finally, we shall argue that the network is more likely to invest in residual rights of less valuable programs than in the rights for more expensive programs, and this investment decision must be incorporated into the empirical model of the next section.

In this paper, we address ourselves to the determinants of the cost of regularly scheduled new prime-time entertainment series. These series appear sporadically throughout the year as each network discards shows which fail to attract a substantial share of the prime-time audience, but most changes are made at regular intervals of approximately 13 weeks. The largest number of changes occur in the fall and winter, and our sample mainly includes programs which appeared for the first time in the fall of the years 1960 through 1965.

---

[20] *Ibid.*, p. 7.

[21] *Ibid.*, p. 19. Emphasis supplied.

Negotiations for a new series may begin as much as one or two years before a program appears on a network, and it is rare for the lag to be less than six months. At this time, a formal contract or a memorandum of agreement will be drawn up between the supplier and the network. In fact, some of these important details may be omitted until a later date as both parties abide by an informal understanding. Therefore, in most cases, the prices to be paid by the network are stipulated considerably in advance of the network's decision to schedule the program or even order a single episode, and the value of the program series must be determined *a priori* without any definite notion of the nature of the competition which the program will face when it is finally exhibited.

Once the program series is actually ordered, the network begins to search for the appropriate prime-time hour for its exhibition. Approximately 60 to 70 percent of the shows in the rival two networks' schedules will be continuations of last season's series, and these rivals will be reluctant to move continuing series regardless of the new series chosen by the network to appear opposite them. Thus, for the most part, each network will choose to replace its series failures under the assumption that rivals will not move most of their more successful shows in response. We have, therefore, the basis for a "Cournot" theory of programming, where each network takes its rivals' choices as given in attempting to maximize its profits.

Our Cournot model of network triopoly is exposited in terms of a single time period within prime time. The inputs for the ith network's program will be denoted t_{ij}, referring to the number of units of the jth input into the ith network's program. For the first network, for example, there exists a "production function" of the form

$$A_1 = A_1(t_{11}, \ldots, t_{1M}; t_{21}, \ldots, t_{2M}; t_{31}, \ldots, t_{3M}),\ j = 1, \ldots, M, \tag{1}$$

where A_1 represents the audience potential of the first network's program. The inputs to this production function include the services of cameras, settings, film, or tape, and a number of other goods as well as the services of all technical and creative persons involved in "producing" an audience through the creation of a given program. The precise relationship between each of these productive services and the audience yield of the program is extremely difficult to specify empirically; therefore, the analysis is delineated in terms of heterogeneous inputs only for pedagogical purposes in this section. A more detailed elaboration of the Cournot model is undertaken in Section 7 below.

The network sells viewers to advertisers at a price, P_A, and therefore faces a net profit (Π) function:

$$\Pi = P_A A_1 - \sum_{j=1}^{M} p_j t_{1j}, \tag{2}$$

where p_j is the market price of the jth input. Its problem is to choose the t_{1j}'s which maximize (2). This requires that for each t_{1j}:

$$\left[P_A + \frac{dP_A}{dA_1} A_1\right]\left[\frac{\partial A_1}{\partial t_{1j}} + \sum_{j=1}^{M}\left(\frac{\partial A_1}{\partial t_{2j}}\cdot\frac{dt_{2j}}{dt_{1j}}\right) + \sum_{j=1}^{M}\left(\frac{\partial A_1}{\partial t_{3j}}\cdot\frac{dt_{3j}}{dt_{1j}}\right)\right] = \left[p_j + \frac{\partial p_j}{\partial t_{1j}} t_{1j}\right]. \tag{3}$$

The first bracket on the left-hand side of (3) is simply marginal revenue (of audience) facing the network. The second bracket has three terms: the first is the marginal product of the t_{1j} and the next two terms are conjectural response terms for the other two networks. In our Cournot model, the conjectural response terms are equal to zero, thus simplifying the algebra and the analysis considerably. The term on the right-hand side is simply the marginal cost of the jth input.

Under Cournot assumptions, the conjectural response terms in (3) disappear, and we are left with

$$\frac{\left[\dfrac{\eta - 1}{\eta}\right]}{\left[\dfrac{\delta_{j+1}}{\delta_j}\right]} P_A \frac{\partial A_1}{\partial t_{1j}} = p_j, \qquad j = 1, \ldots, M, \tag{4}$$

where η represents the price elasticity of advertising demand facing the network and δ_j is the elasticity of supply of the jth factor to the network. We shall simply denote $(\delta_j + 1)/\delta_j$ by k_j hereafter.

Equation (4) is the first-order condition for the employment of each of the M inputs. Solved for each t_j, it represents a variety of a Cournot reaction curve since it is obvious that $\partial A_1/\partial t_{1j}$ varies with the strength of rival networks' fare. Specifically, it is likely that $\partial^2 A_1/\partial t_{1j}\partial t_{2j}$ and $\partial^2 A_1/\partial t_{1j}\partial t_{3j}$ are negative, implying that networks will "counter-program," placing their more expensive programs opposite the rivals' weakest fare and vice versa.

While (4) holds for any type of input, the mechanism by which it is achieved varies across different varieties of program inputs. For nonreproducible talent or "stars", $\partial A_1/\partial t_{1j}$ should be replaced by a discrete term reflecting the increment to audience attributable to the performer. Moreover, given that the network may only choose whether to employ the performer or not, t_j may not assume values other than zero or unity. It is p_j which varies to satisfy (4) in this case, and as long as this price is above the performer's opportunity cost in other media, he will offer his services to program packagers.

For the other inputs, the network varies t_j according to the magnitude of the marginal productivity term, the price of network advertising, and its monopoly and monopsony power in the respective markets. In practice, however, many nontalent inputs to the production process are used in fixed proportions either because of the nature of the production function or union agreements. For instance, the amount of lighting, the number of technicians, the number of feet of film or videotape, the amount of stage space required, the editing of the film or tape, and numerous other production services are not utilized in significantly variable proportions. Nevertheless, the prices of these services are likely to contain considerable economic rent, since bargaining agents may press for higher wages for the suppliers of the services as their marginal revenue products [according to (4)] rise. But these prices should vary with the average profitability of all entertainment fare in a given year. Thus we expect some nontalent costs to be fixed while others vary with the strength of demand for network advertising.

A simple manipulation of (4) allows us to summarize our model by distinguishing between the rents paid to talent and the prices earned by nontalent inputs. If we assume that the first K of the M inputs are talent, total payments to talent for the first network run (C_t) may be written

$$C_t = \frac{\eta - 1}{\eta}\left[\sum_{j=1}^{K} k_j \frac{\partial A_1^e}{\partial t_{1j}} \cdot \frac{1}{A_1^e}\right] P_A A_1^e, \tag{5}$$

where A_1^e is the expected audience of the given program and $(\partial A_1^e/\partial t_{1j})(1/A_1^e)$ is the elasticity of expected audience with respect to the jth talent input. Similarly, nontalent costs (C_{nt}) may be written

$$C_{nt} = C_0 + \frac{\eta - 1}{\eta}\left[\sum_{j=K+1}^{M} k_j \frac{\partial A_1}{\partial t_{1j}} \cdot \frac{t_{1j}}{A_1}\right] P_A \bar{A}_1, \tag{6}$$

where $\bar{A}$ is the average audience for network programs in the current year and C_0 represents fixed costs which do not vary with the net revenues from the sale of entertainment programming.

5. The empirical estimation of program costs

■ The previous two sections have led us to hypothesize that the cost of any program to a network will vary with the anticipated audience of that program, the average revenue yield of programs in network exploitation, and the residual rights obtained by the network. The basic equation to be estimated from available data is of the form

$$C_i = a_0 + a_1(P_A \bar{A}_i) + a_2(P_A A_i^e) + a_3 n_i D_i + a_4 n_i S_i + u_i, \tag{7}$$

where C_i is the cost per exhibition of the ith program, A_i^e is its expected audience, D_i is a dummy variable which is equal to unity if the network obtains the distribution right for syndication of the program and zero otherwise, and S_i is the profit share from syndication obtained by the network. Both S_i and D_i are multiplied by n_i, the proportion of exhibitions represented by first runs, since the value of syndication interests is directly proportional to the number of episodes filmed. A randomly distributed error term with zero mean and constant variance is denoted u_i.

The magnitude of the coefficients of a_1 and a_2 in (7) will depend upon the relative importance of nontalent and talent expenditures in total program costs and the degree of monopoly or monopsony power exercised by the networks in their output and input markets [see equations (5) and (6)]. If either $(\eta - 1)/\eta$ or the k_j differ substantially from unity, the networks will be paying inputs to the production process less than the value of their marginal product, and a_1 and a_2 will be correspondingly smaller than they would be if the networks were perfect competitors in all markets. Similarly, if a_3 and a_4 are found to be smaller than the discounted present value of accounting returns from syndication, we may infer that the networks possess monopsony power in dealing with their suppliers.

☐ **The data.** Data for new regularly scheduled entertainment series on the three networks in the fall seasons of 1960 through 1965 were obtained from the contracts signed by the networks. In all, contracts for 148 program series were available, and of these 148, complete data were available for 121 series.[22]

The cost datum, C_i, is an estimate of the average cost of one half hour of a single episode in its first exposure on the network. In order to obtain this estimate, it is assumed that the network would schedule 52 episodes during a year by purchasing the maximum number of new episodes possible and filling out the year with repeats. Since repeats generally bring audiences of only 70 percent of new episodes, the total cost of 52 weeks of the series is divided by the sum of the number of new episodes plus 0.7 times the number of repeats required.[23]

As an estimate of $P_A \bar{A}_t$, the total estimated network night-time sales to advertisers for each year is utilized. Since the number of networks and the number of prime-time hours per week did not vary from 1960 through 1965, these industry sales serve well as a proxy for the average revenues per hour during each of the six years studied.

The estimate of *expected* revenues per program, $P_A A_t^e$, was obtained by asking each network to provide its initial price for a single commercial minute of advertising in each series during its fall season. An attempt was made to obtain the prices offered by the networks before advertisers began placing their orders. This set of prices provides the earliest recorded estimates of the networks' own expectations concerning the revenue potential of their programs.

Measures of D_i and S_t are taken directly from the contracts with suppliers. D_t is a set equal to unity if the network owns either the domestic distribution right or the foreign right or both. S_i is equal to the proportion of net profits from foreign and domestic syndication granted to the network company by the contract.

---

[22] For 19 of these programs, there was no record in the network files of the initial asking price for commercial minutes. For 8 others, it was impossible to obtain a precise cost estimate.

[23] In addition, an adjustment in C_t is made for network investment in a pilot and for "overage" agreements. It is assumed that any pilot program is utilized as an episode in the series, and the cost of this pilot is included in the cost of a year's run of the series. Overage payments are network rebates of cost overruns which are remitted from the network's syndication profit share Approximately one-third of network programs earn profits from syndication, therefore, the net value of the overage agreement to the supplier is assumed to equal one-third times the limit to overruns specified in the contract, which, in turn, is usually 5 or 10 percent of stipulated program costs

In the empirical analysis below, two additional series are used. A variable indicating the average strength of programs scheduled by the two rival companies at the same hour, A_t^r, is constructed by averaging the percentage of total television homes watching each rival program in the preceding year.[24] If the rival program is new during the current season, its strength is measured by multiplying the average performance of that network's programs by 0.9, an arbitrary weight designed to reflect a lower probability of audience acceptance. Another variable, H, measures the total homes normally watching television during the specific hour of the week during which the program is scheduled.

☐ **Regression results.** Equation (7) was fitted to data for the entire 121-program sample and for a 108-program subsample which excludes all music variety programs. The exclusion of these programs is based upon an *a priori* judgment that the values of a_3 and a_4 are likely to be much smaller for such fare given the rapid obsolescence of popular music, and that relative production costs for this type of programming are quite different from the costs of other series. Both samples were also utilized to fit a variant of (7) in which the time trend, T_t, was substituted for $P_A\bar{A}_t$ in order to determine if the latter was merely capturing the trend of general cost inflation during 1960–1965. The results demonstrate that $P_A\bar{A}_i$ serves as a better regressor in the program cost equation, but the difference is not statistically significant.

The ordinary least-squares estimate of (7) for both samples is satisfactory, with one glaring exception. The coefficient of the syndication profit share variable is negative and significant, suggesting that networks pay less for programs in which they obtain profit shares. See Table 1. This counterintuitive result derives from

TABLE 1

ORDINARY LEAST–SQUARES ESTIMATES OF PROGRAM COSTS (EQUATION 7)

REGRESSION	SAMPLE SIZE	ESTIMATED COEFFICIENTS FOR THE INDEPENDENT VARIABLES*						
		CONSTANT	$P_A\bar{A}$	$P_A A^e$	D	S	T	R^2
(i)	121	12,304 (2.42)	0.21864 (3 82)	0.18467 (5.52)	664 (0.48)	−7474 (2 03)		0.410
(ii)	108	16,898 (4.30)	0 25948 (5 81)	0.09271 (3 35)	2483 1 (2.30)	−6888.3 (2.31)		0 448
(iii)	121	27,590 (8.33)		0.18569 (5 54)	662.44 (0 48)	−7384.5 (2 00)	1128.5 (3 75)	0.408
(iv)	108	35,119 (13 07)		0.09434 (3.39)	2482.4 (2.28)	−6740 7 (2.24)	1339.9 (5.65)	0.440

*t STATISTICS IN PARENTHESES.

MEAN $P_A\bar{A}$ = \$90,000

MEAN $P_A A^e$ = \$98,611 FOR 121–PROGRAM SAMPLE; \$97,965 FOR 108–PROGRAM SAMPLE [BASED UPON 0 85 TIMES THE ASKING PRICE FOR THREE COMMERCIAL MINUTES IN EACH PROGRAM].

SOURCE OF MEAN $P_A\bar{A}$: LITTLE REPORT [3], P. 85.

---

[24] This series is simply last year's Nielsen rating for continuing shows.

the networks' tendency to purchase larger shares in the less expensive programs with lower audience appeal. It is not perfectly clear why this pattern should exist, but an explanation may be attempted. The networks tend to place the weakest of their new programs opposite their rivals' strongest returning series; therefore, program suppliers are less interested in retaining profit shares in these shows, since they are aware that low-cost series are less likely to survive on the networks than more expensive shows. Given that a program series must have a fairly long network run to accumulate sufficient episodes to make syndication profitable, program suppliers may be less interested in retaining syndication interests in any series with relatively low probability success on the network.

It thus becomes necessary to estimate (7) as part of a simultaneous equation system which includes a profit-share investment function. Unfortunately, not many exogenous variables recommend themselves for such an equation, since there is no obvious *a priori* theory which predicts the outcome of negotiations for these subsidiary rights. The following simple linear form is utilized for the purposes of this investigation:

$$S_i = b_0 N_i^a + b_1 N_i^c + b_2 N_i^n + b_3 R_i + b_4 C_i + b_5 T_i + v_i. \quad (8)$$

The N_i variables are dummies for each of the three networks, A.B.C., N.B.C., and C.B.S., respectively. R_i is a dummy variable equal to unity if the program is an hour series and zero if it is a half-hour series. This variable is included simply because the networks tended to procure profit shares in hour series more than in half-hour series during our sample period. A time trend, T_i, is included to capture the response of the networks to increasing FCC concern about network ownership of profit shares during the period.

When (7) is estimated by two-stage least squares utilizing all exogenous variables in (7) and (8) in the first stage, a very large value for a_4, the coefficient of S, results. An examination of the residuals reveals that the relationship is misspecified. The expected value of a program is measured by its anticipated revenues for a given competitive situation, but if networks systematically place their weaker fare against the rivals' strongest programs and vice versa, $P_A \bar{A}^e$ does not measure the value of each program for a given competitive situation. Put another way, this variable does not reflect the value of specialized inputs in attracting audience under invariant conditions. It cannot be expected that a star performer will accept a lower rent in performing in a program which is to be sacrificed before another network's most popular series.

In order to adjust $P_A A^e$ for the effect of network rivalry, this variable is multiplied by the square root of A^r,[25] the average audience appeal of other network programs appearing during the same time period. But including A^r in equation (7) introduces a further complication, since we expect the cost of a program to vary *directly* with the value of $P_A A^e (A^r)^{0.5}$ but we also expect each network to schedule its most expensive programs opposite its rivals' least successful series. Thus we expect the scheduling decision to find C varying *inversely* with A^r. In order to avoid biased estimates of (7), we must introduce equation (9) to reflect the scheduling decision for each program. This scheduling relationship includes each of the network dummy variables as well as H_i, a measure of the number of households normally watching television during the time period in which the ith program is scheduled.

---

[25] Expected revenues are multiplied by the *square root* of the average of rivals' program ratings for the past year because a log-linear regression of *ex post* audience upon program cost, rivals' program ratings from last year, A^r, and television homes watching at the given hour, H, yielded a coefficient for A^r of approximately 0.5. Equation (7′) was also estimated with $R_i^e = (p_A A_i^e)(A_i^r)$, but the results were little different from those in Table 2. The proportion of explained variance fell very slightly, and the value of $a_3 + a_4$ was reduced by approximately 20 percent.

Our complete model may now be written as follows:

$$C_i = a_0 + a_1(P_A\bar{A})_i + a_2R_i^e + a_3n_iD_i + a_4n_iS_i + u_i \qquad (7')$$

$$S_i = b_0N_i^a + b_1N_i^c + b_2N_i^n + b_3R_i + b_4C_i + b_5T_i + v_i \qquad (8)$$

$$C_i = c_0N_i^a + c_1N_i^c + c_2N_i^n + c_3H_i + c_4A_i^r + w_i, \qquad (9)$$

where R_i^e is equal to $(P_A A_i^e)(A_i^r)^{0.5}$ and u_i, v_i, and w_i are random disturbance terms with zero mean and constant variance. There are three endogenous variables, C_i, A_i^r, and S_i, in a simultaneous system of three equations.

Estimates of the parameters of (7′) are obtained by the method of two-stage least squares. Reduced-form estimates of A_i^r and S_i are first obtained by ordinary least-squares regressions of each variable upon all exogenous variables. The reduced form estimates of R_i^e and n_iS_i are then obtained by regressing each on all exogenous variables and the corresponding transformation of each variable's reduced form. The second-stage estimates of (7′) are presented in Table 2, along with approximate t statistics and coefficients of multiple determination.

TABLE 2

TWO–STAGE LEAST–SQUARES ESTIMATE OF PROGRAM COSTS (EQUATION 7′)

REGRESSION	SAMPLE SIZE	ESTIMATED COEFFICIENTS FOR THE INDEPENDENT VARIABLES*					
		CONSTANT	$P_A\bar{A}$	R^e	D	S	R^2
(v)	121	15,923 (2 793)	0 18224 (3 069)	0.16172 (5.710)	1256.9 (0.7627)	741.6 (0.05246)	0 418
(vi)	108	15,900 (3.501)	0.21328 (3.820)	0 10270 (3.297)	1893.8 (1.493)	12,285 (1 412)	0.295

*t STATISTICS IN PARENTHESES

As expected, the regression for the smaller sample is quite different from the one which includes observations on music-variety programs, and the large differences in the values of the coefficients of $P_A\bar{A}$ and R^e support our assertion that the ratio of talent to other production costs varies substantially between music-variety and all other program types. Regression (vi) provides an estimate of the combined value of the distribution right and a 100-percent share of profits from syndication of \$1,894 plus \$12,285 times the mean value of n(0.788), or \$11,031 per new half-hour episode of entertainment programming. The sum of the coefficients of these variables is equal to 1.7 times the standard error of the sum. Expressed as a percentage of mean program costs for the period, these syndication interests amount to 23 percent of network program costs or 19 percent of the total value of network exhibition and syndication interests. These results are compared with published estimates of syndication profits in Section 6.

Of equal interest is the low value of the constant term in regressions (v) and (vi), suggesting that only $16,000 of the costs of an average half-hour program were not price determined during the period. The relative importance of the other two terms can best be explained by evaluating them at the point of means of the regression. The mean value of program costs is $48,725 per half-hour for the larger sample and $48,328 for the smaller sample. The mean value of D and S is approximately 0.22 in each sample. For the smaller sample, therefore, the net cost of the network exhibition alone is $48,328 – 0.22(14,179) = $45,209.

For the entire 108-program sample, the mean value of the networks' expected revenue per half hour ($P_A A^e$) is $97,965. Therefore, $a_2 R^e$ may be estimated at $10,045 at the point of means, or 22 percent of total program costs. Our theory leads us to assign all of these costs to talent expenses, but we have only informal evidence that such an allocation is justified. The Little Report found that by 1964 above-the-line costs had risen to approximately 48 percent of total costs. These expenditures include payments for actors, directors, producers, and other personnel. We have no further data on the division of program costs in commercial broadcasting in the United States, but it is interesting to compare the Little datum with data for a foreign noncommercial network operating in competition with a single commercial network.[26] For 1964–1965, above-the-line costs for this foreign network represented 44 percent of total program costs for entertainment programming. Expenditures for performing artists were 22 percent of total costs, or precisely the same as the value of $a_2 R^e$ in regression (vi) above.

At the midpoint of the sample period, average network revenues per half hour ($P_A \bar{A}$) were approximately $90,000. Thus, at the point of means, $a_1 P_A \bar{A}$ is equal to $19,200 in regression (vi) or 42.5 percent of network program costs. This component of costs represents the price-determined payments to program inputs which do not vary across programs within a single year. Our theory leads us to assign these payments to the price-determined elements of nontalent costs, but we cannot identify with precision the recipients of such payments.

6. The value of syndication interests

■ The regression results presented in Table 2 provide an estimate of the value of distribution rights and profit-share income from syndication of 19 percent of network program outlays in 1960–1965. Can this estimate be reconciled with published data on the actual returns from syndication? A comparison with the Little

---

[26] This network has kindly provided detailed data on its program costs which are utilized in Section 7. At the request of the network, its identity is not revealed in this paper.

Report data provides the only available answer to this problem.

In a 1967 prospectus,[27] Twentieth Century-Fox reported that:

> [c]ommencing with the television series produced in 1964, ... [production costs] are being amortized in the proportion that any year's actual net revenue bears to the estimate of total net revenue expected to be realized from networks and syndication distribution This change ... has resulted in the amortization of approximately 80 per cent of production cost of films released each year against television network revenue.

Fox is among the largest suppliers of prime-time series, and there is no reason to believe that its experience is not representative of the entire industry. As a check on this estimate, it is useful to compare the data in the Little Report on network program payments and total program payments to suppliers and their own production subsidiaries of $290.3 million for prime-time series in 1964, and total "off network" syndication sales of $78 million in the same year.[28] Sales were 27 percent of prime-time program costs in 1964. This may be a misleading comparison, however, since syndication revenues follow network program payments by the networks in the same series with a mean lag of perhaps one or two years. In 1965–1966, the average flow of syndication sales from old network series was $89.4 million, or 31 percent of 1964 program payments by networks.[29] Some of these sales may have come from series not originally shown on the networks during prime time, but it is unlikely that there was a large contribution from this source. It seems reasonable, therefore, to assume that the ratio of syndication sales to networks' prime-time program costs was 0.30 in 1964.

The network data appearing in the Little Report show that new series cost the networks about $125,000 per hour episode in 1964, an average which was rising at an annual rate of approximately 7 percent at the time.[30] Since at least one-third of network prime-time series are new to the network in a typical year, it seems reasonable to assume that the average cost per new episode shown on the networks during prime time in 1964 was $110,000 per hour. But the networks license programs for more than one exhibition. Therefore, we require data on the payments for repeats as well as new exhibitions.

Once more the Little Report provides the necessary data. In 1964, the networks paid $290.3 million for prime-time entertainment series. Given 71½ hours of such programming per week,[31] the average cost per hour may be estimated at $78,000. During 1964–1965, the networks exhibited original episodes for most series from mid-September until mid-April, filling out the remainder of the season with repeat episodes. Thus, original episodes formed 7/12 of the season's offering for an average program, and the ratio of the cost of new episodes to total program payments may be estimated at 7/12 ($110,000/$78,000), or 0.82. Given that syndication revenues are 30 percent of total network program payments, they are 0.30 divided by 0.82 of new program costs (*C*) for the first run of the programs. The ratio of syndication revenues to program costs is thus 0.366. See Figure 2.

---

[27] Dated November 22, 1967, p. 8

[28] See Little Report [3], vol. 2, pp 130 and 141.

[29] *Ibid.*, p. 141.

[30] This datum is used rather than the mean value from the sample collected for the regressions in Section 5 The latter is not a complete census of all new programs. *Ibid.*, p 69.

[31] *Ibid.*, p. 5.

FIGURE 2

CALCULATION OF THE VALUE OF SYNDICATION INTERESTS FROM 1964 COST–REVENUE DATA FOR PRIME-TIME ENTERTAINMENT SERIES

$$\frac{\text{SYNDICATION PROFIT}}{\text{NEW EPISODE COST PER HOUR}} = \left[\frac{\text{PROFIT FROM DISTRIBUTION}}{\text{SYNDICATION SALES}} + \frac{\text{PROFITS FROM PROFIT SHARES}}{\text{SYNDICATION SALES}}\right]\left[\frac{\text{SYNDICATION SALES}}{\text{TOTAL NETWORK PROGRAM PAYMENTS}}\right]\left[\frac{\text{TOTAL NETWORK PROGRAM PAYMENTS}}{\text{TOTAL PROGRAM HOURS}}\right]\left[\frac{\text{TOTAL PROGRAM HOURS}}{\text{TOTAL HOURS OF NEW EPISODES}}\right]\left[\frac{1}{\text{COST PER HOUR OF NEW EPISODE}}\right]$$

$$= \left[0.11 + 0.35\right] \cdot \left[\frac{\$87.1\text{ MILLION}}{\$290\ 3\text{ MILLION}}\right]\left[\frac{\$290.3\text{ MILLION}}{71\tfrac{1}{2} \times 52}\right]\left[\frac{71\tfrac{1}{2} \times 52}{7/12 \times 71\tfrac{1}{2} \times 52}\right]\left[\frac{1}{\$110{,}000}\right]$$

$$= 0.168$$

$$\frac{\text{EXPECTED SYNDICATION PROFIT}}{\text{NEW EPISODE COST PER HOUR}} = \left[1 + \text{EXPECTED RATE OF GROWTH}\right]\left[0.168\right]$$

$$= 1\ 15 \times 0.168 = 0.193$$

The profitability of syndication may also be inferred from the Little Report's interpretation of network data, but only after some adjustment. The networks report a "net gain" from syndication distribution fees of $2,319,000 and net gains from profit-share income of $5,419,000 in 1964.[32] Estimates of producers' (and presumably other participants') shares in the profits from syndication are estimated in a rather simple fashion in the Little Report. The networks' average profit share in the current year's network series is divided into profit-share net gains for 1964 in order to obtain an estimate of the total profits accruing to all participants. Similarly, the calculation of industry profits from distribution fees is effected by simply dividing the network datum by the share of programs in which networks have obtained the rights for currently exhibited network series. Both estimates must be considered to be biased downward.

In the calculation of industry profit-share net gains, the bias occurs in utilizing the net profit figure for the networks. Our data for new program series in the 1960–1965 period reveal that the network profit shares are inversely correlated with program success and, therefore, syndication profitability. As a result, the Little Report technique of dividing network profit share into network imputations of profits from syndication must lead to an underestimate of industry profits from syndication. We cannot know the magnitude of this downward bias, but the Little Report estimate of $19.5 million of industry profits from this source in 1964—35.8 percent of net revenues after deducting distribution fees[33]—

---

[32] *Ibid.*, p. 132

[33] *Ibid.*, p. 130. It is assumed that the average distribution fee is 0.30 and that producers' profit-share income from syndication represents the same share of total profit-share income as for the network companies.

must be adjusted upward. Given that only shipping costs, print costs, production overages, and a few miscellaneous expenses are deducted from net syndication revenues before computing profit shares, it would be surprising if total profit shares did not amount to at least 50 percent of net revenues.

The estimate of industry distribution fees is also too low, as a simple assumption will demonstrate. There is no reason to believe that other distributors are less efficient than the network subsidiaries which realized profit margins of 11.0 percent of sales in 1964. Since industry sales were $78 million, industry profits from distribution fees should approximate $8.58 million, not $6.5 million as estimated by the Little Report.[34]

The network company which was bidding for syndication interests at the midpoint of our sample—the beginning of the 1963 season—could hardly be unaware that syndication profits were growing at a very rapid rate in 1960–1963, a rate of more than 30 percent according to the Little estimates. This company would expect syndication profits to be realized from its successful programs far past 1964; therefore, the simple estimates for 1964 calculated in the above paragraphs are probably considerably below the profits expected by interested parties in 1963. There is no way of knowing what growth rate might have been applied by these bidders, but given an annual rate of growth of syndication sales of 15 to 20 percent[35] and a more rapid rate of growth of profits, it would not seem unreasonable to apply a 15-percent rate to the above profit data.

The final estimate of syndication profitability is derived as follows. Syndication fees contribute 11.0 percent of total sales in net profits while profit-share income adds 0.50 of 70 percent of sales (obtained by deducting 30 percent for average distribution fees). Therefore, 1964 profits from syndication were 0.46 of total sales. Given a growth rate of 0.15 percent, *expected* profits may be assumed to be 0.529 of 1964 *actual* syndication sales. Since syndication sales represent 0.366 of new program costs (C), syndication profits from the distribution fee and a 100 percent profit-share are estimated at 0.529 times 0.366 times new program costs, or 0.193 C for the average program during the midpoint of our sample period. This procedure is depicted schematically in Figure 2.

The final estimate is admittedly rough, relying upon network data in which important cost imputations cannot be known. There is a considerable increase in the estimated profitability of profit shares from that appearing in the Little Report, but all other estimates are quite conservative.

It remains to estimate the *present value* of 0.193 C since syndication revenues are not realized immediately upon payment of program costs, but only after a considerable lag. Since foreign sales may begin immediately and account for nearly 60 percent of all revenues, it seems reasonable to assume that the bulk of syndication revenues may be realized in four years. Again using a conservative estimate, assuming that these revenues are realized evenly over the four-year period and that the opportunity cost of capital in 1963–

---

[34] *Ibid.*

[35] *Ibid.*, p. 141.

1964 was 0.10, the present value of 0.193 C may be calculated as follows:

$$[(0.193\ C)/4] \int_0^4 e^{-rt}dt = 0.04825C[(1 - e^{-0.4})/0.10] = 0.159\ C.$$

A long, tortuous route has brought us to an estimate of the value of syndication profits, which is nearly 16 percent of the price of a new episode of a program series to the networks.

How sensitive are these estimates to the assumptions made along the way? For instance, if the Little data were taken literally, the estimate of the present value of syndication profits would fall to 11.6 percent of new program costs. If all other assumptions were to be retained, an increase in the assumed cost of capital to 15 percent would lower present value from 0.159 to 0.145 C. An assumed growth rate of syndication profits of 20 percent would raise the estimate to 16.6 percent. Only the denial of the assumptions in the Little Report's calculations of industry profits seems of great importance, but this denial seems warranted even though our adjustment is certainly a rough one.

The final accounting estimate of the value of all syndication interests—16 percent of new program costs to the networks—is well within the confidence interval of our econometric estimate of 19 percent (Table 2). It would therefore appear that networks have paid an amount equal to the discounted present value of the distribution rights and profit shares which they have obtained. The Prime Time Rule has not saved the program suppliers from the rapacious tendencies of network buyers; it has simply required the suppliers to assume more risk or to sell these syndication interests to other buyers.

7. An extension of the results

■ If we extend the theoretical model developed earlier in this paper, we may utilize the regression results along with further empirical analysis to make a rough estimate of the degree of monopoly-monopsony power possessed by the networks.

For instance, suppose that (1) is appropriately represented by (1′):

$$A_1 = \frac{f\left(\sum_{j=1}^{M} t_{1j}\right)}{F\left(\sum_{n=1}^{3}\sum_{j=1}^{M} t_{nj}\right)} \qquad \begin{array}{c} \partial A_1/\partial t_{1j} > 0 \\ \partial A_1/\partial t_{nj} < 0 \\ n \neq 1. \end{array} \tag{1′}$$

If we assume further that $f'(\ ) = F'(\ )$, then:

$$\frac{\partial A_1}{\partial t_{1j}} \cdot \frac{t_{1j}}{A_1} = \frac{f_j'(\ )t_{1j}}{f(\ )}\left[1 - \frac{f(\ )}{F(\ )}\right], \tag{10}$$

where $f_j'(\ )t_{1j}/f(\ )$ is the elasticity of the network's own audience with respect to its own investment in the jth input. If we call this elasticity μ_j, we may insert (10) into (4) to obtain

$$\left|\frac{\eta - 1}{\eta}\right| \cdot P_A A_1 \sum_{j=1}^{M} k_j \cdot \mu_j\left[1 - \frac{f(\ )}{F(\ )}\right] = \sum_{j=1}^{M} p_j t_{1j}. \tag{11}$$

At Cournot equilibrium, $1 - f(\ )/F(\ )$ becomes $1 - 1/N$, where N is the number of rival networks. If we assume further that all $k_j = k$, we may write the expression for Cournot equilibrium as

$$k\left[\frac{\eta - 1}{\eta}\right]\left[\frac{N - 1}{N}\right]\left[\sum_{j=1}^{M} \mu_j\right][P_A A_1] = \sum_{j=1}^{M} p_j t_{1j}. \qquad (12)$$

For talent inputs alone, this relationship simply states that above-the-line expenditures are equal to the product of the audience expected from the program and the sum of the elasticities of audience with respect to each input multiplied by three terms which reflect the intensity of competition among networks for audience, advertising revenue, and performers. In our regression results of the previous section, the estimate of a_2 is

$$a_2 = k\left[\frac{\eta - 1}{\eta}\right]\left[\frac{N - 1}{N}\right]\left[\sum_{j=1}^{K} \mu_j\right] \approx 0.10.$$

If we could determine the value of the sum of the elasticities, $\sum_{j=1}^{K} \mu_j$, we would be able to estimate the value of $k[\eta - 1/\eta]$ for $N = 3$, the number of networks throughout our sample period.

Unfortunately, we do not have precise data on the distribution of above-the-line and below-the-line costs for our U. S. sample, but the foreign noncommercial data do permit us to attempt to estimate $\sum_{j=1}^{K} \mu_j$. These foreign data contain observations on above-the-line costs, payments to performing artists, and audience for 54 light entertainment and drama series. The light entertainment series—mostly filmed comedy or light drama and music-variety programs—are likely to be chosen to provide the maximum audience per unit cost, but the drama series may well be chosen to satisfy the tastes of a sizeable minority despite the costs involved in their production. For this reason, empirical investigation into cost-audience relationships are likely to be useful only for the light-entertainment series.

In order to obtain a measure of $\sum_{j=1}^{K} \mu_j$ from the noncommercial program experience, we estimate two forms of simple regressions of audience upon costs—one which is linear and one which is log-linear.[36] The independent variables are either artists' costs or all above-the-line costs and the dependent variable is average audience per program. The results are presented in Table 3 for filmed comedy or light drama series and for music-variety series as well for a combination of the two. The final column in Table 3 contains the estimated elasticity of audience with respect to the independent variable at the point of means. These elasticity measures for all above-line costs range from 0.5 to more than 0.7, and we may take 0.6 as an approximation to the mean estimate.

---

[36] The regression of audience upon either component of costs provides us with a direct estimate of the effect of talent inputs upon audience The linear regression gives us a slope coefficient equal to $\sum_{j=1}^{K} (\partial A/\partial t_j)$ and the loglinear regression yields slope coefficients equal to $\sum_{j=1}^{K} \mu_j$. This result may be proved quite simply. If audience is denoted A and total factor payments of the appropriate category are denoted $\sum_{j=1}^{K} p_j t_j$ and if we assume that $A = \alpha\left[\sum_{j=1}^{K} p_j t_j\right]^{\beta}$, $\partial A/\partial t_j = (\beta A p_j)/\left(\sum_{j=1}^{K} p_j t_j\right)$ and $(\partial A/\partial t_j)(t_j/A) = \beta(p_j t_j)/\left(\sum_{j=1}^{K} p_j t_j\right)$. Summing over all K inputs gives $\sum_{j=1}^{K} (\partial A/\partial t_j)(t_j/A) = \beta$ But $\sum_{j=1}^{K} \mu_j = \sum_{j=1}^{K} (\partial A/\partial t_j)(t_j/A)$; therefore, $\sum_{j=1}^{K} \mu_j = \beta$.

TABLE 3

ELASTICITY OF AUDIENCE WITH RESPECT TO EXPENDITURES UPON TALENT, FOREIGN NONCOMMERCIAL NETWORK, 1964–1965*

	SAMPLE	CONSTANT TERM	COSTS OF PERFORMING ARTISTS	ABOVE–THE–LINE COSTS	R^2	ESTIMATED ELASTICITY
LINEAR FORM REGRESSION COEFFICIENTS	ALL LIGHT ENTERTAINMENT (n = 29)	3341 (3.02)	1 350 (3.77)		0.345	0 513
		3262 (2 73)		0.7202 (3 51)	0 313	0.533
	FILMED COMEDY AND LIGHT DRAMA (n = 13)	4489 (1 62)	1.089 (1 39)		0.150	0.446
		4034 (1 21)		0 5983 (1 28)	0 130	0.503
	MUSIC–VARIETY PROGRAMS (n = 14)	2846 (2.24)	1.535 (3 26)		0.470	0 537
		2577 (1.94)	1 001 (3.29)		0 475	0 591
LOG LINEAR FORM REGRESSION COEFFICIENTS	FILMED COMEDY LIGHT DRAMA (n = 13)	4.201 (1 70)	0.5846 (1.91)		0.248	0 585
		2 525 (0 78)		0 7262 (1 98)	0.262	0.726
	MUSIC–VARIETY PROGRAMS (n = 14)	3.478 (1.82)	0.6740 (2 63)		0.366	0.674
		3.063 (1.36)		0.6807 (2 42)	0 328	0.681

*t STATISTICS IN PARENTHESES.

If networks had no monopoly nor monopsony power, $k(\eta - 1)/\eta$ would equal unity. Therefore, if $\sum_{j=1}^{M} \mu_j$ is equal to 0.6, we would expect $[(N - 1)/N] \sum_{j=1}^{M} \mu_j$ to equal 0.4 for a three-network oligopoly. Our econometric estimate of this expression (a_2) is only 0.1; therefore, application of the Cournot model leads us to reject the hypothesis of competition in both markets. While we cannot dismiss the possibility of network monopsony in some input markets, our results in Sections 5 and 6 provide some evidence that k is not likely to differ much from unity. These results suggest that $\eta - 1/\eta$ is approximately equal to 0.25 and that η is equal to 1.33 for each network.

8. Summary

If there is a single theme to this paper it is that offhand attempts to connect the fewness in television network broadcasting to theories of monopoly or monopsony behavior are likely to be unsuccessful. Network broadcasting is far from a typical economic activity since its principal output is donated to viewers while advertisers purchase rights to address these viewers with commercial messages. Networks may have monopoly power in national advertising markets, but this market power is not easily identified in output markets since it may only be implemented through a restriction in the number of viewer hours offered to advertisers in any period. It is obvious that this restriction of output is not serious among the current network triopoly, for each firm broadcasts during every hour possible and more than 60 percent of the nation's homes watch this output during prime viewing hours. It is only the FCC—through its imposition of the Prime Time Rule—which has sought to restrict network output to advertisers and viewers alike by limiting prime-time network broadcasting to three hours per day.

Suppose that each of the three networks faces a market demand curve which is much less than infinitely price elastic. Our rather crude estimate of this price elasticity is -1.33 for each firm, an estimate which is certainly inconsistent with the hypothesis of perfect competition in national advertising markets. But what power does this confer upon the networks? Surprisingly, virtually the only effect of this imperfectly competitive situation is to allocate the rents from television broadcasting to the networks and their stations rather than to television celebrities. The size of the total prime-time audience is likely to be rather insensitive to small changes in program quality, and performers are available to the networks in inelastic supply; therefore, the effect of monopoly power in advertising markets is to lower the derived demand for talent used in programs without much change in the nature of those programs. Star performers available at $20,000 per week are also available at a mere $10,000 per week. In our Cournot (competitive) model of program expenditure, monopoly power in advertising markets is simply translated into lower rents for performers.

Obviously, the impact of network monopoly in advertising markets upon program quality depends upon the relative importance of the various types of resources used in producing programs. Our econometric estimates of program costs suggest that as much as two-thirds of all program outlays are directly dependent upon the revenue potential of network programs in their advertising market. We cannot be sure how much of this two-thirds share represents economic rents paid to participants in program production, but fragmentary evidence on the division of program expenditures suggests that a very large share of these price-determined outlays is appropriated by performers and other talent. If there were greater competition in national advertising markets, the derived demand for these services would rise and be reflected in larger economic rents. Profits would be diverted from television networks and their stations to television celebrities.

The FCC has not connected the networks' position in their output markets with their programming behavior, but rather it has focused upon the networks' possession of *monopsony* power. Its restrictions upon network investment in syndication interests is based upon the assumption that the networks do not pay their suppliers the full value of these syndication rights, an assumption which is not confirmed by the econometric estimates presented in Section 5, above. Comparison of these estimates with accounting returns from syndication markets demonstrates that the networks purchase syndication interests at prices which reflect the present value of net revenues generated in syndication markets. This conclusion should surprise no one outside of the Commission, for it is inconceivable that the networks could force suppliers to accept a rate of return on program investments which is below the opportunity cost of capital. If monopsony power were present, it would be reflected in a reduced quantity of programs ordered for network exhibition, not in exploitation and eventual bankruptcy of program suppliers. Most firms supplying network entertainment series during the past decade are alive and well and still competing for network business.

The results of this paper should not be a discouragement to the FCC. That the Commission is not required to protect the small suppliers of entertainment programs—to a large extent, the major motion picture companies—from the rapacious tendencies of the television networks should be a source of considerable relief. Moreover, the identification of the relationship between network profits and market power in national advertising markets should cause the Commission no concern. The major change which would emanate from the development of greater competition in advertising markets would be a redistribution of profits toward television performers. Clearly, the FCC prefers to see these profits in the hands of the networks and stations where they may be devoted to "public service" programming. Of course, the Commission could consider the possibility of taking the necessary steps to permit an increase in the number of national networks, but this would require a reduction in the subsidies paid to local programs which attract few viewers and a major increase in the popular fare so deprecated by the Commission and its staff.

References

1. CRANDALL, R. W. "The Economic Effect of Television-Network Program 'Ownership'." *Journal of Law and Economics,* Vol. 14, No. 2 (October 1971), pp. 385–412.
2. FRIEDLAENDER, A. F. *The Dilemma of Freight Transport Regulation* Washington, D. C The Brookings Institution, 1969.
3 LITTLE, A D., Incorporated. "Television Program Production, Procurement, Distribution, and Scheduling." Unpublished manuscript 1969. Referred to in the article as "Little Report."
4 MACAVOY, P. W. "The Effectiveness of the Federal Power Commission." *The Bell Journal of Economics and Management Science,* Vol. 1, No. 2 (Autumn 1970), pp 271–303.
5. ROBBINS, G. B. "A Study of the Network Base Hourly Rate." Unpublished bachelor's thesis. Massachusetts Institute of Technology, 1971.
6. U. S. FEDERAL COMMUNICATIONS COMMISSION. *Final Cable Television Decision.* F.C.C. Docket 18397 et al., 1972.
7. ———. *Interim Reports: Program Procurement.* F.C.C. Docket 12782, 1960 and 1965.
8. ———. *Notice of Proposed Rule-Making.* F.C.C. Docket 12782, 1965.
9 ———. *Report and Order.* F C.C. Docket 12782, 1970.

Perfect competition and optimal production decisions under uncertainty

Eugene F. Fama
Professor of Finance
University of Chicago

This paper studies production decisions by firms in a capital market where decision-making by consumer-investors and the structure of equilibrium expected returns on securities are according to a two-parameter (mean-dispersion) model. An attempt is made to distinguish clearly questions concerned with the conditions required for perfect competition among firms as issuers of securities from questions concerned with whether value maximizing production decisions by perfectly competitive firms are Pareto optimal. The two major findings are as follows. First, there are natural externalities in the production decisions of firms that could in principle prevent the capital market from being perfectly competitive. But it is argued that the available empirical evidence is consistent with the conditions required for perfect competition. Second, given a market that satisfies the conditions required for perfect competition, natural external economies in the production decisions of firms can prevent the allocations achieved by perfectly competitive, value maximizing firms from being Pareto optimal. That is, a competitive equilibrium need not be Pareto optimal.

1. Introduction

■ Recent work in finance has been concerned with developing the "two-parameter" portfolio models of Markowitz, Tobin, and Fama[1] into theories of capital market equilibrium. These models ask: What can be said about equilibrium prices for investment securities in a market in which consumers make portfolio decisions according to the two-parameter model? The initial work on this "capital asset pricing model" attacks this question within the context of a market in which supplies of securities and production decisions of firms are assumed to be fixed.[2] An exception is Mossin[3] who considers the nature of optimal production decisions under the assumptions that (1) the capital market is perfectly competitive; and (2) in their production decisions, firms try to maximize the market value of common stock outstanding before the decision is

---

Eugene F. Fama received the B.A. degree from Tufts University (1960) and the Ph.D. degree from the University of Chicago (1964). He has been a professor at Chicago since 1968. Professor Fama has published numerous articles in financial analysis, and is coauthor of a recent book on valuation theory.

Research for this paper was supported by a grant from the National Science Foundation. The comments of Professors Fischer Black, Michael Jensen, Dale Jorgenson, Arthur Laffer, John B. Long, Jr., Merton Miller, and William Sharpe are gratefully acknowledged.

[1] See [15], [20], and [6], respectively.

[2] See, for example, Sharpe [21], Lintner [14], and Fama [7].

[3] [17].

made—an objective henceforth referred to as "the market value rule."

Recently, however, the work of Jensen and Long and Stiglitz[4] has raised somewhat related questions concerning: (1) whether there are realistic circumstances in which firms can be perfect competitors in the capital market of the two-parameter model; and (2) whether in such a market, production decisions by firms, acting according to the "market value rule," are Pareto optimal.[5] This paper is concerned with answering these questions.

Although they suggest that firms are not perfectly competitive in the capital market of the two-parameter model, neither Jensen and Long nor Stiglitz attempt to analyze what perfect competition requires in this model. Rather they concentrate directly on whether the allocation of resources that results from the production decisions of value maximizing firms is Pareto optimal—concluding that in general it is not. But it is important to know whether any deviation from Pareto optimality is due to a violation of the conditions required for perfect competition or whether it is the result of something, like an externality, that could be present even in a perfectly competitive market. We argue that because they do not investigate the nature of perfect competition, the Jensen-Long and Stiglitz analyses cannot effectively identify the sources of suboptimality.

This paper attempts to distinguish clearly questions concerned with the conditions required for perfect competition in the capital market of the two-parameter model from questions concerned with whether value maximizing production decisions by perfectly competitive firms are Pareto optimal. The two major findings are as follows. First, there are natural externalities in the production decisions of firms that could in principle prevent the capital market from being perfectly competitive. But it is argued that the available empirical evidence is consistent with the conditions required for perfect competition. Second, given a market that satisfies the conditions required for perfect competition, there can be natural external economies in the production decisions of firms that can prevent the allocations achieved by perfectly competitive, value maximizing firms from being Pareto optimal. That is, a competitive equilibrium need not be Pareto optimal.

2. The two-period, two-parameter model

■ Attention is limited to a discrete time, two-period world in which there is a single resource or good w available at the beginning of period 1, and w can either be consumed at period 1, in which case it is called consumption c_1, or used at period 1 to produce consumption c_2 to be available at period 2.[6] Various techniques of production are assumed to be available: in simplest terms, these techniques are the ways in which w can be transformed at period 1 into probability distributions of c_2. The probability distributions of c_2 produced by firms are assumed to be jointly normal, although the analysis is also applicable to the more general symmetric stable or "stable Paretian" two-parameter model.[7] It is also assumed that there are riskless production techniques; with these the amount of c_2 to be obtained at period 2 from any given input of w at period 1 is known for certain at period 1.

---

[4] See [12] and [23], respectively.

[5] A Pareto optimal set of period 1 consumption-portfolio decisions by individuals and production decisions by firms implies that there is no other set of feasible decisions such that at least one person has higher expected utility and no other person has lower expected utility

[6] The phrase "at period t" is meant to imply "at the beginning of period t."

The resource w is assumed to be owned by consumers, and all production activities are undertaken by firms. Firms "finance" their period 1 production activities by selling shares in their period 2 outputs to consumers in exchange for some of the resource w.[8] Given an equilibrium set of prices for shares, the problem facing a consumer at period 1 is to optimally split his holding of w between current consumption and investment in some portfolio of shares. The return on this portfolio provides his consumption for period 2. All consumers are assumed to be risk averse and to make their consumption-investment decisions according to the criterion of maximum expected utility. Joint normality of the probability distributions of c_2 produced by all firms then implies that the distributions of c_2 (or equivalently, the distributions of portfolio return) available to any consumer are normal, so that the consumer can choose among these distributions by comparing only their means and standard deviations. It is also assumed that there is complete agreement among consumers concerning the joint distribution of period 2 outputs implied by any given set of production decisions by firms at period 1.

A firm is assumed to make a production decision at period 1 that maximizes the difference between the current market value (in terms of the resource w) of the distribution of its period 2 output and the amount of the resource required at period 1 to generate this distribution of period 2 output. This is equivalent to maximizing the current value of the shares outstanding before the production decision is made. Thus production decisions are assumed to be made according to the market value rule.

If we also assume a "frictionless" capital market[9] and the availability to consumers of riskless borrowing and lending opportunities, then for any given set of production decisions by firms, the market value P_j of firm j at the beginning of period 1 is[10]

$$P_j = \frac{1}{1+r}\left[\bar{V}_j - \lambda\,\frac{\operatorname{cov}(\tilde{V}_j, \tilde{V}_m)}{\sigma(\tilde{V}_m)}\right], \tag{1}$$

where tildes ($\tilde{}$) are used to denote random variables, and where

$\bar{V}_j$ = the mean of the distribution of $\tilde{V}_j$, the output of firm j at period 2;
$\operatorname{cov}(\tilde{V}_j, \tilde{V}_m)$ = the covariance between $\tilde{V}_j$ and $\tilde{V}_m = \sum_k \tilde{V}_k$, the aggregate period 2 output of all risky firms;

---

[7] See Fama [7]. If distributions of returns on all portfolios are assumed to be normal (or symmetric stable), then distributions of asset returns must be jointly normal (or symmetric stable). For a discussion of multivariate stable distributions, see Press [18].

[8] In referring to period 2 quantities, the terms consumption, output, and value can and will be used interchangeably

[9] That is, individual consumers are price-takers and there are no transactions costs in exchanging the resource w for shares in firms or in exchanging the shares themselves.

[10] Cf. Sharpe [21], Lintner [14], and Fama [7].

$\bar{V}_m$ and $\sigma(\tilde{V}_m)$ = the mean and standard deviation of the distribution of $\tilde{V}_m$;
$1/(1 + r)$ = the price at period 1 of a unit of expected period 2 output. Since $1/(1 + r)$ is also the period 1 price of a unit of output to be obtained for certain at period 2, r is the "riskless rate of interest" for both borrowing and lending; and
$-\lambda/(1 + r) < 0$ = the period 1 market price per unit of $\text{cov}(\tilde{V}_j, \tilde{V}_m)/\sigma(\tilde{V}_m)$. We argue below that the latter quantity can be interpreted as the risk of the period 2 output of firm j.

The period 1 capital market prices $1/(1 + r)$ and $-\lambda/(1 + r)$ are in units of the resource w, or equivalently, current consumption c_1, which is always the numeraire. And the total period 2 risky output $\tilde{V}_m = \sum_k \tilde{V}_k$ is, of course, also the aggregate period 2 consumption generated by the risky industry.

Next note that

$$\sigma(\tilde{V}_m) = \sum_j \sum_k \text{cov}(\tilde{V}_j, \tilde{V}_k)/\sigma(\tilde{V}_m) = \sum_j \text{cov}(\tilde{V}_j, \tilde{V}_m)/\sigma(\tilde{V}_m). \qquad (2)$$

Thus the period 1 total market value of all risky firms, $P_m = \sum_k P_k$, is

$$P_m = \frac{1}{1 + r}[\bar{V}_m - \lambda\sigma(\tilde{V}_m)]. \qquad (3)$$

Equations (1)-(3) illustrate how, in the two-parameter model, two dimensions—the mean and standard deviation (or "risk") of the distribution of total period 2 output (or consumption)—are priced in the capital market at period 1. The period 1 price of a unit of mean is $1/(1 + r) > 0$, so that mean is a desirable commodity. The price of a unit of standard deviation is $-\lambda/(1 + r) < 0$, so that, in line with the assumption that consumers are risk averse, standard deviation or risk is an undesirable commodity. Moreover, in equation (1) the period 1 market value of the uncertain period 2 output of firm j is simply the sum of the market values of: (1) the contribution, $\bar{V}_j$, of the firm's period 1 production decision to the total expected output of all firms, and (2) $\text{cov}(\tilde{V}_j, \tilde{V}_m)/\sigma(\tilde{V}_m)$, which, from equation (2), is the contribution of the firm's production decision to the risk or standard deviation of aggregate period 2 output.

Although we have assumed that the capital market is perfect in the sense of being frictionless, we are initially concerned with the conditions under which the market is also perfectly competitive in the sense that the period 1 production decisions of an individual firm have no effect on the capital market prices $1/(1 + r)$ and $-\lambda/(1 + r)$, from which the prices of the shares of all firms are derived. The competitiveness of the capital market is analyzed under the assumption that firms make production decisions according to the market value rule. Then the Pareto optimality of the market value rule in a competitive market is discussed. At each stage, comparisons with the results of Jensen and Long and Stiglitz[11] are made.

---

[11] In [12] and [23], respectively

3. Production decisions and perfect competition in the two-parameter model

■ **The perfect certainty analogue.** The first step is to see what can be learned about the problem of perfect competition with a finite number of firms from the treatment of similar problems in perfect certainty models of general equilibrium. Specifically, consider a multi-factor, multi-industry world of perfect certainty in which the following conditions hold:

(C1) Factors of production are infinitely divisible and fully and costlessly mobile across firms and industries. Moreover, either there are no factors that are used in only one industry, or such factors are owned by individuals and their services are sold to firms.
(C2) Information about the returns earned by factors in different firms and industries is available without cost to everybody, and factors act to maximize returns.
(C3) The same techniques of production are available to all firms in any given industry.
(C4) The demand curve for the good produced by any industry is downward sloping with nonzero elasticity at all levels of output.
(C5) At any given level of output, the curve showing the minimum marginal cost of industry output as a function of industry output has greater slope than the industry demand curve.
(C6) Conditions of production are such that the rest of an industry can always use precisely the incremental quantities of factors of production demanded or released by a firm to offset precisely output changes by that firm.

Conditions (C1)-(C3) are standard. It is well to note, however, that they are meant to rule out restrictions on entry. Conditions (C4) and (C5) rule out the types of "unusual" demand or cost conditions that could lead to problems with the stability of an industry equilibrium.[12] Finally, for our purposes, the major implication of condition (C6) is that offsetting output changes by the rest of an industry in response to an output change by an individual firm must also precisely offset any externalities associated with the firm's new decision.

In Fama and Laffer[13] it is shown that under conditions (C1)-(C6) if there are two or more noncolluding firms in each industry, a general equilibrium is perfectly competitive in the sense that individual firms face perfectly elastic demand curves for their outputs. The outline of the argument is as follows. Under conditions (C1)-(C6), a general equilibrium implies that returns to any factor are equal across all firms and industries. In order to maintain this equality of returns, other firms must respond by offsetting any output changes by an individual firm. It is further shown that under conditions (C1)-(C6), other firms respond by precisely offsetting output changes by an individual firm, so that the decisions of a firm have no effect on prices or on industry outputs.[14]

---

[12] If the industry is perfectly competitive, then (C5) implies that at any given level of output, the industry supply curve has greater slope than the industry demand curve, so that equilibrium is unique. But (C1)-(C6) are meant to be a minimum set of conditions that imply perfect competition, which in turn implies that the minimum marginal cost of industry output as a function of industry output is also the industry's supply curve. Thus, since the concern is with minimum conditions that imply perfect competition, (C5) is not stated directly in terms of supply curves.

[13] [9].

This "reaction principle" of offsetting output changes by other firms in response to output changes by an individual firm is the fundamental mechanism whereby the decisions of an individual firm have no effect on price or on industry output. Given appropriate definitions of output, the reaction principle is the essence of perfect competition in any certainty or uncertainty model. We see later, however, that the principle is violated in the "competitive equilibrium" analyses of Jensen and Long and Stiglitz.

☐ **Application to the two-parameter model: preliminary discussion.** The fact that there is uncertainty about the outcomes to be obtained at period 2 from the production decisions made at period 1 in itself causes no difficulty in applying the general equilibrium analysis of Fama and Laffer to the two-parameter model. In the two-parameter model there are three goods in the market at period 1: (1) current consumption, c_1, which is also the input for producing future consumption, (2) expected future output or consumption for period 2, and (3) contributions to the standard deviation or risk of the distribution of the total period 2 output of all firms. Likewise, in the market at period 1 there are the prices $1/(1 + r)$ and $-\lambda/(1 + r)$ for future expected output and contributions to total market risk, both prices stated in terms of current consumption. And although in the risky industry there is uncertainty at period 1 about realized output at period 2, for any given set of production decisions by firms at period 1, the contributions of firms to total expected period 2 output and to the standard deviation of total period 2 output (the things being priced in the market at period 1) are perfectly certain at period 1.

Thus it would seem that the two-parameter model just proposes a specific view of the market context within which a general equilibrium must be established, but that in principle the results of Fama and Laffer apply. That is, if conditions (C1)-(C6) hold for both the riskless and the risky industries, then the capital market is perfectly competitive in the sense that the period 1 production decisions of an individual firm have no effect on the capital market prices $1/(1 + r)$ and $-\lambda/(1 + r)$ and no effect on aggregate outputs of expected period 2 consumption and risk.

Conditions (C1)-(C5) are no more restrictive in the two-parameter model than in the perfect certainty model. And, of course, to assume that condition (C6) holds for the riskless industry in the two-parameter model is no different than to assume that it holds in the certainty model. In fact, the analysis of perfect competition in the two-parameter model is only of special interest with respect to applying condition (C6) to the risky industry. Specifically, the output decisions of individual firms in the risky industry can lead to a novel type of externality: when a firm expands or contracts its output of risk, this in general changes the risk outputs of other firms, even when these other firms do not explicitly change their output decisions. The first step is to illustrate the nature of these externalities. Then we examine what seems to be a not-so-special case where even with externalities the capital market can be perfectly competitive.

---

[14] The offsetting output changes may come either from existing firms, including the possibility of exit, or from entry of new firms.

☐ **The problem: externalities and correlated outputs.** For current purposes it is convenient to measure the period 2 output of a firm in units of its expected period 2 output. That is, define

$$\tilde{v}_j = \frac{\tilde{V}_j}{\bar{V}_j}.$$

Thus v_j is the period 2 output of firm j per unit of its expected output. In terms of this new variable,

$$\sigma^2(\tilde{V}_m) = \sum_j \sum_k \bar{V}_j \bar{V}_k \operatorname{cov}(\tilde{v}_j, \tilde{v}_k) \tag{4}$$

and

$$\operatorname{cov}(\tilde{V}_j, \tilde{V}_m) = \sum_k \operatorname{cov}(\tilde{V}_j, \tilde{V}_k) = \bar{V}_j \sum_k \bar{V}_k \operatorname{cov}(\tilde{v}_j, \tilde{v}_k). \tag{5}$$

For simplicity assume also that any firm in the risky industry produces expected value and covariances in fixed proportions. That is, the different production decisions that any firm j might make at period 1 are such that for all k, $\operatorname{cov}(\tilde{v}_j, \tilde{v}_k)$ is the same for any scale of production chosen, that is, for any levels of $\bar{V}_j$ and $\bar{V}_k$. In short, different production decisions represent different scales of a firm's activities, but per unit of expected value, the distributions of period 2 output associated with different production decisions are the same and relate in the same ways to the distributions of the period 2 outputs (per unit of expected value) of other firms.

Without loss of generality, suppose now that there are only two noncolluding firms in the risky industry. From (1), (3), (4), and (5), in a general equilibrium, the period 1 market values of these firms and the total market value of all risky firms are, respectively,

$$P_1 = \frac{1}{1+r}\{\bar{V}_1 - \lambda[\bar{V}_1 \bar{V}_2 \operatorname{cov}(\tilde{v}_1, \tilde{v}_2) + \bar{V}_1^2 \sigma^2(\tilde{v}_1)]/\sigma(\tilde{V}_m)\},$$

$$P_2 = \frac{1}{1+r}\{\bar{V}_2 - \lambda[\bar{V}_1 \bar{V}_2 \operatorname{cov}(\tilde{v}_1, \tilde{v}_2) + \bar{V}_2^2 \sigma^2(\tilde{v}_2)]/\sigma(\tilde{V}_m)\},$$

and

$$P_m = P_1 + P_2 = \frac{1}{1+r}\{(\bar{V}_1 + \bar{V}_2) - \lambda[2\bar{V}_1 \bar{V}_2 \operatorname{cov}(\tilde{v}_1, \tilde{v}_2)$$
$$+ \bar{V}_1^2 \sigma^2(\tilde{v}_1) + \bar{V}_2^2 \sigma^2(\tilde{v}_2)]/\sigma(\tilde{V}_m)\}.$$

Given such an initial general equilibrium at period 1, suppose firm 1 decides to expand production, that is, increase $\bar{V}_1$ by one unit. If firm 2 does not respond, the output of risk of firm 2 nevertheless changes from

$$[\bar{V}_1 \bar{V}_2 \operatorname{cov}(\tilde{v}_1, \tilde{v}_2) + \bar{V}_2^2 \sigma^2(\tilde{v}_2)]/\sigma(\tilde{V}_m)$$

to

$$[(\bar{V}_1 + 1)\bar{V}_2 \operatorname{cov}(\tilde{v}_1, \tilde{v}_2) + \bar{V}_2^2 \sigma^2(\tilde{v}_2)]/\sigma(\tilde{V}_m)',$$

where

$$\sigma(\tilde{V}_m) = [\bar{V}_1^2\sigma^2(\tilde{v}_1) + \bar{V}_2^2\sigma^2(\tilde{v}_2) + 2\bar{V}_1\bar{V}_2 \operatorname{cov}(\tilde{v}_1,\tilde{v}_2)]^{1/2}$$

and

$$\sigma(\tilde{V}_m)' = [(\bar{V}_1 + 1)^2\sigma^2(\tilde{v}_1) + \bar{V}_2^2\sigma^2(\tilde{v}_2) + 2(\bar{V}_1 + 1)\bar{V}_2 \operatorname{cov}(\tilde{v}_1,\tilde{v}_2)]^{1/2}.$$

The induced change in the output of risk of firm 2 is not simply due to the fact that in general the new aggregate output of risk, $\sigma(\tilde{V}_m)'$, is not equal to the old, $\sigma(\tilde{V}_m)$. In addition to this effect of the output change of firm 1 on the output of risk of firm 2, the numerator of the expression for the output of risk of firm 2 rises if $\operatorname{cov}(\tilde{v}_1,\tilde{v}_2) > 0$ or falls if $\operatorname{cov}(\tilde{v}_1,\tilde{v}_2) < 0$.[15] In short, a change in the output of one firm in the risky industry in itself changes the output of risk of the other firm, and this is the type of natural externality that exists in the two-parameter model.[16]

But for the analysis of the conditions required for perfect competition, the important point is that even if $\sigma^2(\tilde{v}_1) = \sigma^2(\tilde{v}_2)$ and even if firm 2 contracts $\bar{V}_2$ by one unit by releasing precisely the quantity of the resource that firm 1 needs to expand $\bar{V}_1$ by one unit, these seemingly offsetting actions need not return the market to its initial equilibrium position. In general

$$(\bar{V}_1 + 1)(\bar{V}_2 - 1) \operatorname{cov}(\tilde{v}_1,\tilde{v}_2) \neq \bar{V}_1\bar{V}_2 \operatorname{cov}(\tilde{v}_1,\tilde{v}_2),$$

so that even though the firms may seem to be offsetting each other's production decisions,

$$[(\bar{V}_1 + 1)^2\sigma^2(\tilde{v}_1) + (\bar{V}_2 - 1)^2\sigma^2(\tilde{v}_2) + 2(\bar{V}_1 + 1)(\bar{V}_2 - 1) \operatorname{cov}(\tilde{v}_1,\tilde{v}_2)]^{1/2} \neq \sigma(\tilde{V}_m);$$

that is, the resulting quantity of total variance of period 2 market value is not the same as it was in the initial position. In short, in the two-parameter model with correlated period 2 market values, there can be natural externalities that violate condition (C6) of a perfectly competitive market.[17]

☐ **Perfect competition and linear models.** Condition (6) can hold, however, if it is possible to express the period 2 output of any firm as a linear function of random variables whose realized values affect the outputs of subsets of firms, plus a residual whose realized value is independent of the outputs of other firms.[18] As in the classical linear regression model, the residual is assumed to be independent of the "independent variables," but the latter need not be mutually independent. And all random variables are assumed to have multivariate normal (or symmetric stable) distributions. The assumption that the period 2 outputs of firms are multivariate normal (or multivariate symmetric stable) implies that for any given set of production decisions by firms, period 2 outputs can be expressed in terms of such a linear model. But here it is also assumed that the independent variables are the same for different production decisions. Although this seems a slight additional restriction, it is not necessarily implied by multivariate normality.

---

[15] From the evidence of Blume [2] and others, negative covariances seem to be empirically irrelevant. Thus it is reasonable to interpret the phenomenon as an external diseconomy.

[16] The existence of these externalities was first pointed out by Jensen and Long [12]

[17] Although in the two-parameter model problems with condition (C6) are generally due to correlated period 2 market values, in the example of the text there is a problem even when $\operatorname{cov}(\tilde{v}_1,\tilde{v}_2) = 0$. This additional problem arises from the fact that $\sigma^2(\tilde{V}_1) = \bar{V}_1^2\sigma^2(\tilde{v}_1)$ goes up in proportion to the square of expected period 2 output: in effect, all units of the firm's output are perfectly correlated. But if, when $\operatorname{cov}(\tilde{v}_1,\tilde{v}_2) = 0$, $\sigma^2(\tilde{V}_1) = \bar{V}_1\sigma^2(\tilde{v}_1)$ so that in effect units of output are uncorrelated within the firm as well as between the two firms, then the problem disappears.

[18] In the statistical literature these residuals would be called disturbances. In the current context, however, the term residual is more descriptive

To illustrate what the assumed linear model for period 2 outputs implies for the analysis of perfect competition, consider a simple special case, the so-called market model suggested originally in Markowitz, developed further in Sharpe and Fama, and tested empirically in Fama et al., Blume, and other papers.[19] In this model, as applied in the context of this paper, it is hypothesized that the period 2 output of any firm j can be expressed as a linear function of a normally distributed market factor $\tilde{M}$, plus a normally distributed random residual $\tilde{\epsilon}_j$, which is assumed to have zero mean and to be independent of $\tilde{M}$ and of the residuals in the period 2 outputs of other firms. Thus for all j

$$\tilde{V}_j = A_j + B_j\tilde{M} + \tilde{\epsilon}_j, \tag{6a}$$

$$\tilde{V}_m = \sum_j \tilde{V}_j = \sum_j A_j + \tilde{M}\sum_j B_j + \sum_j \tilde{\epsilon}_j, \tag{6b}$$

$$\sigma^2(\tilde{V}_m) = (\sum_j B_j)^2\sigma^2(\tilde{M}) + \sum_j \sigma^2(\tilde{\epsilon}_j). \tag{6c}$$

In this model all dependence among the period 2 outputs of firms arises from the fact that all are related to the market factor $\tilde{M}$. Indeed, that part of a firm's period 2 output that is associated with the market factor is perfectly correlated with that part of any other firm's period 2 output that is associated with the market factor.[20] On the other hand, that part of a firm's period 2 output that is due to the residual $\tilde{\epsilon}_j$ is uncorrelated with the period 2 outputs of all other firms. And these properties of the market model are, of course, reflected in the expression for $\sigma^2(\tilde{V}_m)$.

We now see that the combination of perfectly correlated and uncorrelated components proposed in this model is consistent with condition (C6) and thus with a perfectly competitive capital market. Using equations (6a) and (6b), first note that in the market model

$$\operatorname{cov}(\tilde{V}_j, \tilde{V}_m) = B_j(\sum_k B_k)\sigma^2(\tilde{M}) + \sigma^2(\epsilon_j), \tag{7}$$

so that, from equations (1), (3), and (6),

$$P_j = \frac{1}{1+r}\{\bar{V}_j - \lambda[B_j(\sum_k B_k)\sigma^2(\tilde{M}) + \sigma^2(\tilde{\epsilon}_j)]/\sigma(\tilde{V}_m)\} \tag{8}$$

and

$$P_m = \frac{1}{1+r}\{\bar{V}_m - \lambda[(\sum_j B_j)^2\sigma^2(\tilde{M}) + \sum_j \sigma^2(\tilde{\epsilon}_j)]/\sigma(\tilde{V}_m)\}. \tag{9}$$

---

[19] See [15], [22], [6], [11], and [2], respectively.

[20] It is important to note that, more generally, in a multifactor model, that part of a firm's period 2 output that is associated with a particular factor is perfectly correlated with that part of any other firm's output that is associated with that factor.

In this model a change in the production decision of firm j is reflected in changes in some combination of the values of $\bar{V}_j$, B_j and $\sigma^2(\tilde{\epsilon}_j)$. Aside from the effects of output changes on aggregate risk $\sigma^2(\tilde{V}_m)$, all externalities in the production decisions of firm j center on changes in the level of B_j. Thus, from (7), if there are no offsetting reactions by other firms to an increase in the level of B_j, the resulting increase in $\sum_k B_k$ directly affects the output of risk of all other firms, i, such that $B_i \neq 0$. On the other hand, and again aside from the implied changes in $\sigma(\tilde{V}_m)$, changes in the level of $\sigma^2(\tilde{\epsilon}_j)$ have no such effect on the risk outputs of other firms. The reason, of course, is that $\tilde{\epsilon}_j$ is uncorrelated with the period 2 outputs of all other firms, whereas $B_j\tilde{M}$ is perfectly correlated with the market factor components in the outputs of all other firms.

Inspection of equations (6c), (8), and (9) reveals that if a change by firm j in the chosen level of B_j is matched by an equal but offsetting change by another firm or firms, then the latter change also precisely offsets the externalities implied by the change in the production decision of firm j. That is, $\sum_k B_k$ remains the same so that firms that do not change their output decisions experience no changes in their outputs of risk, and the total standard deviation of the period 2 outputs of all firms is unchanged. Likewise, if changes by firm j in the chosen levels of $\bar{V}_j$ and $\sigma^2(\tilde{\epsilon}_j)$ are matched by equal but offsetting changes by another firm or firms, then firms that do not change their output decisions experience no changes in their outputs of mean or of risk, and the mean and standard deviation of the sum of the period 2 outputs of all firms are unchanged.

Thus condition (C6) holds in the market model if other firms can precisely offset changes in $\bar{V}_j$, B_j, and $\sigma^2(\tilde{\epsilon}_j)$ by any individual firm j with precisely the quantities of the resource demanded or released by firm j.[21] Then since we have already assumed conditions (C1)-(C5), the results of Fama and Laffer imply that the reaction principle of a perfectly competitive market holds: Given an initial general equilibrium, the equilibrium response of other firms to any change in the production decisions of an individual firm is precisely offsetting. As a result, the total supplies of period 2 mean and standard deviation and the equilibrium capital market prices $1/(1 + r)$ and $-\lambda/(1 + r)$ are unaffected by the production changes of any individual firm, so that the capital market is perfectly competitive.

---

[21] As discussed in Fama and Laffer [9], in either a certainty or uncertainty model, literal satisfaction of condition (C6) for every possible output change by an individual firm requires production functions characterized by constant returns to scale. With constant returns, although the output of an industry in a general equilibrium is determinate, the output or size of an individual firm is not. But aside from some annoying "discontinuities," the Fama-Laffer analysis also holds when there are nonconstant returns to scale—more specifically, when there is a finite optimal firm or plant size. Then one only assumes that condition (C6) holds for output changes by individual firms that are multiples of the optimal firm or plant size. For such output changes, it is again true that price per unit of output is unaffected by the production decisions of individual firms. In other cases, the general equilibrium response by other firms to a change in output by an individual firm may be to offset only partially the firm's output change, though the ultimate response must be sufficient to make it unprofitable for another whole firm or plant to enter or leave the industry. One is, of course, hard pressed to reconcile the assumption of infinitely divisible factors with the existence of a finite optimal firm or plant size. But we do not wish to get into that issue here

The analysis can be extended directly both to the case where there are multiple normally distributed factors that are related to the period 2 outputs of firms and to the case where the period 2 outputs of firms can be expressed as linear functions of stable nonnormal variables.

Linear models of period 2 outputs would seem to be a short additional step from the assumption that period 2 outputs of firms are multivariate normal (or multivariate symmetric stable). More important, however, the available empirical evidence[22] indicates that such linear models do a good job in describing actual return data. Indeed, even the oversimplified market model has been found to be useful in a wide variety of empirical applications.

☐ **Comparisons with Jensen-Long and Stiglitz.** Although they do not analyze the conditions required for perfect competition, Jensen and Long and Stiglitz state that the capital market of the two-parameter model is not perfectly competitive. Jensen and Long imply that perfect competition requires that outputs are perfectly correlated across firms.[23] But the preceding analysis shows that perfect competition is possible as long as the period 2 outputs of firms can be expressed as linear functions of components that are perfectly correlated across firms and components that are uncorrelated across firms.

On the other hand, Stiglitz suggests that the capital market of the two-parameter model is not perfectly competitive because, under his definition of size, the market value of a firm increases less than in proportion with its size.[24] The basis of his argument is easily illustrated in the case where outputs are uncorrelated across firms, so that the valuation equation (1) reduces to[25]

$$P_j = \frac{1}{1+r}\left[\bar{V}_j - \lambda \frac{\sigma^2(\tilde{V}_j)}{\sigma(\tilde{V}_m)}\right].$$

From this equation it is easy to see that if one firm produces twice the mean and twice the standard deviation of another firm, the market value of the larger firm is less than twice the market value of the smaller. But it is also clear that if the mean and variance of one firm is twice that of another, the market value of the larger firm is twice that of the smaller. Thus the question is whether, following Stiglitz, the size of a firm's risk output should be measured in terms of standard deviation, or whether, as the valuation equation seems to suggest, size is appropriately measured in terms of variance. We argue that the latter view is correct.

---

[22] See, for example, Blume [2], Fama et al. [11], or King [13].

[23] "The individual firm in our model, to the extent that it has access to new investment projects whose returns are not perfectly correlated with the returns of projects available to other firms, enjoys some monopoly power in the sale of claims to its new projects" [12], p. 173

[24] ". . . although the market for the trading of securities is probably one of the most competitive markets in existence, the firms do not act as perfect competitors in issuing securities; they are aware that . increasing their size will increase their market value less than in proportion" [23], p. 55.

[25] That is, in this case,

$$\operatorname{cov}(\tilde{V}_j, \tilde{V}_m) = \sum_k \operatorname{cov}(\tilde{V}_j, \tilde{V}_k) = \sigma^2(\tilde{V}_j).$$

Suppose that to produce twice the mean and standard deviation of period 2 output, the large firm also requires twice as much period 1 input of the resource as the smaller firm. Since the market value of the larger firm is less than twice that of the smaller, the market's signal is that the larger firm should not exist; and indeed in equilibrium the only firm sizes that can exist are those that provide the maximum difference between market value and input (which maximum must always be zero). Moreover, from the viewpoint of economic efficiency, the signal provided by market values is correct: the larger firm contributes the same amount to aggregate expected period 2 output as two smaller firms, but it contributes more to the standard deviation of aggregate period 2 output than the two smaller firms.

On the other hand, a firm that produces twice the mean and variance of period 2 output as another firm, contributes exactly twice as much to both the mean and standard deviation of period 2 output. Given that the larger firm requires twice the period 1 input of the resource as the smaller firm, from the viewpoint of economic efficiency it is a matter of indifference as to whether production is carried out by larger or smaller firms. And that is precisely the signal provided by market values.

In short, Stiglitz' conclusion that the capital market is not competitive results from using the wrong units to measure a firm's output of risk. When, as is directly implied by equations (1)–(3), a firm's outputs are measured in terms of its contributions to the mean and standard deviation of aggregate period 2 output, then, as in any perfectly competitive market, values are always proportional to output.

The problems discussed above are errors of interpretation in the Jensen-Long and Stiglitz analyses that have little effect on their substantive results with respect to the Pareto optimality of value maximizing production decisions by firms. But their failure to consider directly the nature of perfect competition in the two-parameter model leads to more serious difficulties. Their analyses uniformly violate the fundamental reaction principle of a perfectly competitive market. All of the Stiglitz model and much of the Jensen-Long model is concerned with the case where in making its production decision, the firm assumes that the production decisions of other firms are fixed. As a consequence, the firm takes account of the effects of its decision on aggregate outputs. This is essentially the approach of the Cournot model of monopolistic competition.[26] And if one starts with a model of monopolistic competition, it is hardly surprising that the resulting allocation of resources to production is not Pareto optimal.

Thus the Jensen-Long and Stiglitz conclusion that elements of monopolistic competition are endemic to the capital market of the two-parameter model is in my opinion false. Their methodology *imposes* elements of monopolistic competition on their results. They are misled by a "mixed mode" line of reasoning in which any individual firm is assumed to be a price-taker in the capital market, but at the same time changes in its production decisions are allowed to affect aggregate output—and it takes account of these effects. From the viewpoint of a competitive model, many "anomalies" follow from this approach, but they are less anomolous when one realizes that the basic methodology is not consistent with a competitive model. Two examples suffice.

---

[26] See, for example, Fama and Laffer [9]

First, note that if the market is perfectly competitive, then in the valuation equation (1), r, λ, and $\sigma(\tilde{V}_m)$ are independent of the production decisions of an individual firm, and an equivalent way of writing equation (1) is

$$P_j = \frac{1}{1+r}[\bar{V}_j - \lambda' \operatorname{cov}(\tilde{V}_j, \tilde{V}_m)], \tag{1a}$$

where $\lambda' = \lambda/\sigma(\tilde{V}_m)$. In their analysis of Pareto optimality, however, Jensen and Long consider the case where λ' is assumed to be independent of a firm's production decisions separately from the case where λ is so fixed, obtaining somewhat different results. But the differences arise from the fact that although firms are assumed to be price-takers with respect to either r and λ in equation (1) or r and λ' in equation (1a), a change in a firm's production decision is allowed to change $\sigma(\tilde{V}_m)$ and the firm takes this into account.

Next, suppose again that the market is perfectly competitive, and, for simplicity, that the linear mechanism generating the period 2 outputs of firms is the market model. Then in the valuation equation (8), a unit change in B_j leads to the same change in the market value of firm j, irrespective of the level of B_j. This follows from the reaction principle of perfect competition whereby production changes by an individual firm have no effect on r, λ, $\sigma(\tilde{V}_m)$, and $(\sum_k B_k)\sigma^2(\tilde{M})$. But if, following Stiglitz, one assumes instead that r, $\lambda' = \lambda/\sigma(\tilde{V}_m)$, and the production decisions of other firms are fixed, then again from equation (8), the change in market value associated with a unit change in B_j is not independent of the level of B_j.[27] From here it is a short step to the conclusion that the value to a firm of an investment with given stochastic return properties depends on the size of the firm, so that the investment would be evaluated differently by different value maximizing firms. This turns out to imply a Pareto suboptimal allocation of resources. But when one properly takes account of the reaction principle of perfect competition, the result disappears.

4. Competitive equilibrium and Pareto optimality

■ Since the Jensen-Long and Stiglitz analyses do not deal with a fully competitive model, one cannot tell whether any absence of Pareto optimality in the production decisions of firms arises from the monopolistic elements that they impose on the capital market, or whether such absence is endemic to the two-parameter world. The final question to be addressed here is whether the equilibrium that results when the capital market of the two-param-

---

[27] Mathematically, in the competitive model,

$$\partial B_j(\sum_k B_k)\sigma^2(\tilde{M})/\partial B_j = (\sum_k B_k)\sigma^2(\tilde{M}),$$

whereas in the Stiglitz model,

$$\partial B_j(\sum_k B_k)\sigma^2(\tilde{M})/\partial B_j = (\sum_k B_k)\sigma^2(\tilde{M}) + B_j\sigma^2(\tilde{M})$$

eter model is perfectly competitive is also Pareto optimal. We find that because of what can be interpreted as natural external economies of scale in the production of uncorrelated risky outputs, in general, a competitive equilibrium is not Pareto optimal—that is, the production decisions of value maximizing, perfectly competitive firms are not Pareto optimal.

Begin by looking at the market model in a slightly different way.[28] In particular, suppose the expected period 2 output of a firm in the risky industry is expressed as $\bar{V}(B,\sigma^2(\tilde{\varepsilon}))$. That is, in order to generate $\bar{V}(B,\sigma^2(\tilde{\varepsilon}))$ units of expected period 2 output, the firm must also generate at least B units of market factor risk and $\sigma^2(\tilde{\varepsilon})$ units of residual variance. Without loss of generality, assume also that the market factor $\tilde{M}$ is scaled so that $E(\tilde{M}) = 0$. Thus the model can be written

$$\tilde{V} = A + B\tilde{M} + \tilde{\varepsilon} = \bar{V}(B,\sigma^2(\tilde{\varepsilon})) + B\tilde{M} + \tilde{\varepsilon}.$$

Let $I(B,\sigma^2(\tilde{\varepsilon}))$ be the minimum quantity of the resource that a firm in the risky industry needs to generate $\bar{V}(B,\sigma^2(\tilde{\varepsilon}))$ and assume for simplicity that all firms in the industry use the same technology,[29] that is, they choose the same values of B and $\sigma^2(\tilde{\varepsilon})$. Then with the current version of the market model, a firm's output of risk—the contribution of its output to $\sigma(\tilde{V}_m)$—is[30]

$$\frac{\operatorname{cov}(\tilde{V},\tilde{V}_m)}{\sigma(\tilde{V}_m)} = \frac{B(NB)\sigma^2(\tilde{M}) + \sigma^2(\tilde{\varepsilon})}{\sigma(\tilde{V}_m)}, \tag{10}$$

where N is the number of firms in the risky industry. Thus the period 1 market value of a risky firm can be written as[31]

$$P = \frac{1}{1+r}\left\{\bar{V}(B,\sigma^2(\tilde{\varepsilon})) - \lambda\frac{B(NB)\sigma^2(\tilde{M}) + \sigma^2(\tilde{\varepsilon})}{\sigma(\tilde{V}_m)}\right\}. \tag{11}$$

The value maximizing, perfectly competitive firm tries to maximize $P - I(B,\sigma^2(\tilde{\varepsilon}))$. But a competitive equilibrium implies that the maximum value of $P - I(B,\sigma^2(\tilde{\varepsilon}))$ is zero, and that at the value maximizing outputs of B and $\sigma^2(\tilde{\varepsilon})$, the marginal value of $P - I(B,\sigma^2(\tilde{\varepsilon}))$ with respect to changes in either B or $\sigma^2(\tilde{\varepsilon})$ is zero. Formally, using equation (11),

$$\bar{V}(B,\sigma^2(\tilde{\varepsilon})) - (1+r)I(B,\sigma^2(\tilde{\varepsilon})) - \lambda\frac{B(NB)\sigma^2(\tilde{M}) + \sigma^2(\tilde{\varepsilon})}{\sigma(\tilde{V}_m)} = 0, \tag{12}$$

$$\frac{\partial\bar{V}(B,\sigma^2(\tilde{\varepsilon}))}{\partial B} - (1+r)\frac{\partial I(B,\sigma^2(\tilde{\varepsilon}))}{\partial B} - \lambda\frac{(NB)\sigma^2(\tilde{M})}{\sigma(\tilde{V}_m)} = 0, \tag{13}$$

$$\frac{\partial\bar{V}(B,\sigma^2(\tilde{\varepsilon}))}{\partial\sigma^2(\tilde{\varepsilon})} - (1+r)\frac{\partial I(B,\sigma^2(\tilde{\varepsilon}))}{\partial\sigma^2(\tilde{\varepsilon})} - \lambda\frac{1}{\sigma(\tilde{V}_m)} = 0. \tag{14}$$

---

[28] Since in this section the capital market is assumed to be perfectly competitive, the period 2 outputs of firms in the risky industry are assumed to be generated by some linear mechanism like the market model. As in the previous section, it is easy to show that the results to be obtained with the market model are easily generalized to multifactor linear models.

[29] Thus, for the moment, the notation can be simplified by dropping subscripts to indicate individual firms

[30] Cf. equation (7)

[31] Cf. equation (8)

These equations make use of the fact that in a perfectly competitive market, a firm's production decisions have no effect on λ and $(1 + r)$ and no effect on the aggregate quantities supplied, NB and $\sigma(\tilde{V}_m)$. Thus, in determining how much B to generate, the perfectly competitive firm takes $(1 + r)$ and $\lambda(NB)\sigma^2(\tilde{M})/\sigma(\tilde{V}_m)$ as given, while in determining how much $\sigma^2(\tilde{\epsilon}_j)$ to generate, the firm takes $(1 + r)$ and $\lambda/\sigma(\tilde{V}_m)$ as given. Alternatively, for those concerned with the problem of observability, simply assume that in the market, the firm only sees the composite price $\theta_B = -\lambda(NB)\sigma^2(\tilde{M})/\sigma(\tilde{V}_m)$, per unit of its output B of market factor risk, and the composite price $\theta_\epsilon = -\lambda/\sigma(\tilde{V}_m)$, per unit of its output $\sigma^2(\tilde{\epsilon}_j)$ of residual variance.

It is easy to show that in the two-parameter model, for each consumer, λ is the marginal rate of substitution (m.r.s.) between standard deviation of period 2 consumption and the expected "risk premium" of period 2 consumption; the risk premium being the difference between expected consumption from the chosen risky portfolio and the consumption that could be obtained by investing all portfolio funds at the riskless rate.[32] It follows directly from equations (12)-(14) that for each firm, λ is also the marginal rate of transformation (m.r.t.) between risk output and expected risk premium. Thus from equation (12) we get

$$\lambda = \frac{\bar{V}(B,\sigma^2(\tilde{\epsilon})) - (1 + r)I(B,\sigma^2(\tilde{\epsilon}))}{\dfrac{B(NB)\sigma^2(\tilde{M}) + \sigma^2(\tilde{\epsilon})}{\sigma(\tilde{V}_m)}} = \frac{\bar{V}(B,\sigma^2(\tilde{\epsilon})) - (1 + r)I(B,\sigma^2(\tilde{\epsilon}))}{\dfrac{\operatorname{cov}(\tilde{V},\tilde{V}_m)}{\sigma(\tilde{V}_m)}}, \tag{15}$$

while from equation (13),

$$\lambda = \frac{\dfrac{\partial \bar{V}(B,\sigma^2(\tilde{\epsilon}))}{\partial B} - (1 + r)\dfrac{\partial I(B,\sigma^2(\tilde{\epsilon}))}{\partial B}}{\dfrac{(NB)\sigma^2(\tilde{M})}{\sigma(\tilde{V}_m)}} = \frac{\dfrac{\partial \bar{V}(B,\sigma^2(\tilde{\epsilon}))}{\partial B} - (1 + r)\dfrac{\partial I(B,\sigma^2(\tilde{\epsilon}))}{\partial B}}{\partial\left(\dfrac{\operatorname{cov}(\tilde{V},\tilde{V}_m)}{\sigma(\tilde{V}_m)}\right)\Big/\partial B}, \tag{16}$$

and from equation (14),

$$\lambda = \frac{\dfrac{\partial \bar{V}(B,\sigma^2(\tilde{\epsilon}))}{\partial \sigma^2(\tilde{\epsilon})} - (1 + r)\dfrac{\partial I(B,\sigma^2(\tilde{\epsilon}))}{\partial \sigma^2(\tilde{\epsilon})}}{\dfrac{1}{\sigma(\tilde{V}_m)}}$$

---

[32] See, for example, Fama and Miller [10], Chapters 6 and 7. Plotting $E(\tilde{c}_2)$ on the vertical axis and $\sigma(\tilde{c}_2)$ on the horizontal axis, λ is the slope of the consumer's (positively sloping and convex) indifference curve at the point corresponding to the optimal portfolio decision Given riskless borrowing and lending, λ is, of course, also the slope of the boundary of the portfolio opportunity set, i.e., it is the slope of the "efficient set."

$$= \frac{\dfrac{\partial \bar{V}(B,\sigma^2(\tilde{\epsilon}))}{\partial \sigma^2(\tilde{\epsilon})} - (1+r)\dfrac{\partial I(B,\sigma^2(\tilde{\epsilon}))}{\partial \sigma^2(\tilde{\epsilon})}}{\partial\left(\dfrac{\operatorname{cov}(\tilde{V},\tilde{V}_m)}{\sigma(\tilde{V}_m)}\right)\Big/\partial \sigma^2(\tilde{\epsilon})}. \tag{17}$$

From (15), for any firm, λ is the average rate of transformation between risk output and expected risk premium for the firm as a whole, and thus it is the m.r.t. for a new firm contemplating entry at the optimal size. From (16) and (17), λ is also the perfectly competitive firm's m.r.t. with respect to changes in B and $\sigma^2(\tilde{\epsilon})$.

But Pareto optimality also requires that consumer m.r.s.'s are equal to the m.r.t. for the economy as a whole. That is, for the competitive equilibrium to be Pareto optimal, the common marginal rate at which consumers are willing to exchange expected risk premium for risk output must be equal to the marginal rate at which the economy as a whole can produce increased expected risk premium in exchange for increased output of risk. With respect to output of residual variance, this Pareto optimality condition is not met by a competitive equilibrium.

To see this, first note that total period 1 investment, total expected period 2 output, and total risk of the period 2 output of the risky industry are, respectively,

$$I_m = NI(B,\sigma^2(\tilde{\epsilon})), \tag{18}$$

$$\bar{V}_m = N\bar{V}(B,\sigma^2(\tilde{\epsilon})), \tag{19}$$

and

$$\sigma(\tilde{V}_m) = [(NB)^2\sigma^2(\tilde{M}) + N\sigma^2(\tilde{\epsilon})]^{1/2}. \tag{20}$$

With respect to entry by a new firm, for the economy as a whole the relevant m.r.t., call it M.R.T.$_{\cdot N}$, is

$$\text{M.R.T.}_{\cdot N} = \frac{\dfrac{\partial \bar{V}_m}{\partial N} - (1+r)\dfrac{\partial I_m}{\partial N}}{\dfrac{\partial \sigma(\tilde{V}_m)}{\partial N}}$$

$$= \frac{\bar{V}(B,\sigma^2(\tilde{\epsilon})) - (1+r)I(B,\sigma^2(\tilde{\epsilon}))}{\dfrac{2B(NB)\sigma^2(\tilde{M}) + \sigma^2(\tilde{\epsilon})}{2\sigma(\tilde{V}_m)}}. \tag{21}$$

Comparing equations (21) and (15), we see that if $\sigma^2(\tilde{\epsilon}) > 0$, M.R.T.$_{\cdot N} > \lambda =$ m.r.s. The fact that $\sigma^2(\tilde{\epsilon})$ is indeed the source of the problem can be determined by looking at the m.r.t.'s with respect to changes in B and $\sigma^2(\tilde{\epsilon})$ for the economy as a whole:

$$\text{M.R.T.}_{\cdot B} = \frac{\dfrac{\partial \bar{V}_m}{\partial B} - (1+r)\dfrac{\partial I_m}{\partial B}}{\dfrac{\partial \sigma(\tilde{V}_m)}{\partial B}} \tag{22}$$

$$= \frac{\dfrac{\partial \overline{V}(B,\sigma^2(\tilde{\epsilon}))}{\partial B} - (1+r)\dfrac{\partial I(B,\sigma^2(\tilde{\epsilon}))}{\partial B}}{\dfrac{(NB)\sigma^2(\tilde{M})}{\sigma(\tilde{V}_m)}},$$

$$\text{M.R.T.}_{\sigma} = \frac{\dfrac{\partial \overline{V}_m}{\partial \sigma^2(\tilde{\epsilon})} - (1+r)\dfrac{\partial I_m}{\partial \sigma^2(\tilde{\epsilon})}}{\dfrac{\partial \sigma(\tilde{V}_m)}{\partial \sigma^2(\tilde{\epsilon})}}$$

$$= \frac{\dfrac{\partial \overline{V}(B,\sigma^2(\tilde{\epsilon}))}{\partial \sigma^2(\tilde{\epsilon})} - (1+r)\dfrac{\partial I(B,\sigma^2(\tilde{\epsilon}))}{\partial \sigma^2(\tilde{\epsilon})}}{\dfrac{1}{2\sigma(\tilde{V}_m)}}. \tag{23}$$

Comparison of equations (22) and (16) shows that $\text{M.R.T.}_B = \lambda$. But comparison of (23) and (17) shows that $\text{M.R.T.}_{\sigma} > \lambda$; that is, by increasing output of residual variance, the economy as a whole can obtain increased expected risk premium in exchange for increased risk at a rate greater than λ. Since in a competitive equilibrium λ is the corresponding m.r.s. for each consumer, the competitive equilibrium is not Pareto optimal: consumers would prefer a larger output of residual variance than that produced by value maximizing perfectly competitive firms.[33]

Jensen and Long and Stiglitz also obtain the result that value maximizing production decisions by individual firms may not be Pareto optimal.[34] They interpret the result as due in part to the existence of monopolistic competition in the risky industry.[35] But that explanation is not relevant here since the analysis in this section has assumed perfect competition. The result is that the competitive equilibrium is not Pareto optimal. Stiglitz concludes that the problem is also due to correlation of period 2 outputs across firms.[36] But the preceding analysis suggests that in fact the difficulties arise in the production of uncorrelated residual outputs.

---

[33] Some details of this result are of at least secondary interest. It is easy to show that if production of $\sigma^2(\tilde{\epsilon})$ always implies some production of B but the two are not produced in fixed proportion, then in a competitive equilibrium; firms in general choose a value of $\sigma^2(\tilde{\epsilon})$ that is too small relative to the Pareto optimal value, and the number of firms in the risky industry is also too small. If B and $\sigma^2(\tilde{\epsilon})$ must be produced in fixed proportion, then competitive equilibrium implies firms of Pareto optimal size, but there are too few of them. Likewise, if there are two risky industries, one for the production of B and one for the production of $\sigma^2(\tilde{\epsilon})$, then firms in both industries choose optimal sizes; the number of firms in the B industry is Pareto optimal, but there are too few firms in the $\sigma^2(\tilde{\epsilon})$ industry.

[34] In [12] and [13], respectively

[35] "One can argue that the nonoptimality which we have discussed so far is due to the fact that our firms have monopoly power in the capital markets" [12], p. 168

[36] "It seems clear that the result of the model, that the market economy will not be Pareto optimal, both because firms fail to take adequate account of covariances with other firms and because they perceive that their value will not increase in proportion to their scale in the short run, suggests that the stock market may not be as efficient a mechanism for allocating resources as has until now been thought to be the case and, in particular, as has been suggested by the previously formulated models of the stock market" [23], p. 55.

We now argue that the suboptimality of a competitive equilibrium is due more specifically to what can be interpreted as external economies of scale in the production of residual variance that are not taken into account by value maximizing perfectly competitive firms. The nature of these externalities—and the role of the law of large numbers—is best illustrated by looking at two special cases.

First, consider an economy where all risks are residual risks; that is, the period 2 outputs of all firms are independent. Let us revert to using the subscript j to indicate firm j, and write $\sigma^2(\tilde{\epsilon}_j) = q_j\sigma^2(\tilde{\eta})$, so that

$$\sigma(\tilde{V}_m) = [\sum_k q_k\sigma^2(\tilde{\eta})]^{1/2} = \sqrt{q}\,\sigma(\tilde{\eta}),$$

where $q = \sum_k q_k$ is total "standard units of variance," $\sigma^2(\tilde{\eta})$, produced by all firms. In short, what we observe in this expression is the familiar result that the standard deviation of a sum of independent variables only goes up with the square root of the number of variables in the sum—the law of large numbers at work. In this example, the firm's production of risk, the contribution of its output to aggregate risk, $\sigma(\tilde{V}_m)$, is $\text{cov}(\tilde{V}_j, \tilde{V}_m)/\sigma(\tilde{V}_m) = q_j\sigma^2(\tilde{\eta})/\sigma(\tilde{V}_m)$. In a perfectly competitive market, the firm takes $\sigma(\tilde{V}_m)$ as fixed,[37] so that it sees its output of risk changing with q_j at the rate

$$\frac{\partial[\text{cov}(\tilde{V}_j, \tilde{V}_m)/\sigma(\tilde{V}_m)]}{q_j} = \frac{\sigma^2(\tilde{\eta})}{\sigma(\tilde{V}_m)} = \frac{\sigma(\tilde{\eta})}{\sqrt{q}}. \tag{24}$$

But for the economy as a whole, the rate of change of $\sigma(\tilde{V}_m)$ is

$$\frac{\partial\sigma(\tilde{V}_m)}{\partial q_j} = \frac{\sigma^2(\tilde{\eta})}{2\sigma(\tilde{V}_m)} = \frac{\sigma(\tilde{\eta})}{2\sqrt{q}}. \tag{25}$$

In taking total output of risk, $\sigma(\tilde{V}_m)$, as fixed, the value maximizing, perfectly competitive firm neglects the "diversifying effects" or external economies[38] that arise from the fact that the period 2 realizations of residual outputs are independent. The firm, however, is not responding "incorrectly" to its environment. In a perfectly competitive market, $\sigma(\tilde{V}_m)$ is indeed unaffected by its production decisions. But, the diversifying effects of residual outputs can only be observed by looking at the rate of change of $\sigma(\tilde{V}_m)$ with respect to changes in aggregate residual outputs, which the firm has no incentive to do. And comparing equations (17) and (23) with (24) and (25) shows that this analysis is not special to the simple economy considered here.

Next consider an economy where all risks are market factor risks; that is, the period 2 outputs of all firms are perfectly correlated, so that

---

[37] Recall again that the essence of a perfectly competitive capital market is that changes in the firm's production decisions are precisely offset by changes on the part of other firms. Thus, a firm's production decisions have no effect on aggregate supplies and prices of any commodities.

[38] These diversifying effects are external economies in the sense that they apply to output of risk, which is an undesirable commodity.

$$\sigma(\tilde{V}_m) = [(\sum_k B_k)^2\sigma^2(\tilde{M})]^{1/2} = \sum_k B_k\sigma(\tilde{M}).$$

Here there is no law of large numbers at work: the standard deviation of total output goes up directly with $\sum_k B_k$, and there are no diversifying effects. This is, of course, due to the fact that now the period 2 outputs of firms are perfectly correlated. In this example, the contribution of the output of firm j to total risk $\sigma(\tilde{V}_m)$ is $\mathrm{cov}(\tilde{V}_j, \tilde{V}_m)/\sigma(\tilde{V}_m) = B_j(\sum_k B_k)\sigma^2(\tilde{M})/\sigma(\tilde{V}_m)$. In a perfectly competitive market the firm implicitly takes both $\sum_k B_k$ and $\sigma(\tilde{V}_m)$ as fixed, so that

$$\frac{\partial[\mathrm{cov}(\tilde{V}_j, \tilde{V}_m)/\sigma(\tilde{V}_m)]}{\partial B_j} = \frac{(\sum_k B_k)\sigma^2(\tilde{M})}{\sigma(\tilde{V}_m)} = \sigma(\tilde{M}) = \frac{\partial\sigma(\tilde{V}_m)}{\partial B_j};$$

i.e., since there are no external economies in the production of market factor risk, the rate of change of the firm's output of risk is the same as the rate of change for the economy as a whole. And from equations (16) and (22) we can see again that the result is not special to the simple economy considered here. It is because of this result that the decisions of perfectly competitive firms with respect to output of market factor risk can be Pareto optimal.

In short, there is a diversifying effect in the production of residual risk that cannot be observed by competitive, value maximizing firms and which leads to the result that on the production side a competitive equilibrium is not Pareto optimal.[39]

It is important to emphasize, though, that the problem with a competitive equilibrium arises entirely from the components of outputs of individual firms that are uncorrelated with the outputs of all other firms. For example, it is easy to show that in the general multifactor model, in which the period 2 output of a firm is represented as a linear function of normal (or symmetric stable) variables, it is only the residual component of a firm's output—that is, the part that is uncorrelated with the outputs of all other firms—that leads to differences between Pareto optimal and competitive equilibrium production decisions. All other components of the firm's period 2 output, which in the nature of the model are perfectly correlated with the corresponding components in the outputs of at least some other firm, cause no problems.[40]

---

[39] Stiglitz comes close to this result in the statement:

"If the only risk the firm is 'selling' is associated with a common market factor *and if the individual firm is small relative to the economy as a whole,* we obtain the usual competitive results on Pareto optimality. When the firm is 'selling' in addition some risk not associated with the market it places too much emphasis on this" [23], p. 50 (the italics are mine).

The italicized qualification about size of firm is an error that arises from Stiglitz' failure to take account of the reaction principle of a perfectly competitive market (cf. note 27 above and the discussion leading to it). Moreover, the statement above is difficult to reconcile with the following, which seems to attribute Pareto optimality problems to the market factor:

"The reason for this misallocation is that the covariance of the firms with one another acts very much like an externality" [23], p 46.

This last interpretation is the one Stiglitz chooses to emphasize in his concluding remarks See, for example, the quotation in note 36 above.

Some consolation might be obtained from the empirical evidence that the total output of residual variance is in fact trivial relative to the total output of market factor risk. Specifically, the empirical evidence of Blume and King[41] indicates that in the post-World War II period, the market factor accounts for about one-third of the variance in the market value of the average firm's common stock. That is, in the expression

$$\sigma^2(\tilde{V}_j) = B_j^2\sigma^2(\tilde{M}) + \sigma^2(\tilde{\epsilon}_j),$$

$\sigma_j^2(\tilde{\epsilon})$ is about twice as large as $B_j^2\sigma^2(\tilde{M})$. Nevertheless, because of the diversifying effects of residual output, aggregate residual variance is trivial relative to aggregate market factor variance. In particular, if (without loss of generality) we scale $\tilde{M}$ so that the average B_j is equal to one, then in the expression

$$\sigma^2(\tilde{V}_m) = (\sum_k B_k)^2\sigma^2(\tilde{M}) + \sum_k \sigma^2(\tilde{\epsilon}_j),$$

the market factor term is on the order of $N/2$ times the residual variance term, where N, the total number of firms, is of course quite large.

In the interests of rigor, however, some of the usual qualifications from welfare economics must be added to this analysis of the empirical evidence. One is interested in the loss in welfare that results from suboptimal output of residual variance, and although we may suspect that the loss is small, the fact that aggregate residual risk is trivial relative to other sources of risk says nothing directly about this. Moreover, it is impossible to make specific statements about the extent of resource misallocation that results from suboptimal output of residual variance without specific assumptions concerning the terms on which the economy can increase its output of independent risk. For example, it would be necessary to say something rather arbitrary about the degree of jointness in the production of market factor and residual risk.

5. Summary and conclusions

■ The preceding analysis can be summarized as follows. When the period 2 outputs of firms are correlated, there are natural externalities in their period 1 production decisions: an increase in the production of risk by one firm changes the risk outputs of other firms, even though these other firms do not explicitly change their production decisions. But when the period 2 outputs of firms can be represented by some linear process like the market model, then offsetting production changes by other firms in response to a change by an individual firm also precisely offset the externalities associated with the firm's change. If these offsetting changes by other firms can take place with precisely the incremental input of the resource demanded or released by the firm, conditions (C1)-(C6) can hold and the capital market can be perfectly competitive in the sense that changes in the production decisions of individual firms have no effect on the prices $1/(1 + r)$ and $-\lambda/(1 + r)$ and on aggregate production of expected period 2 consumption and risk.

---

[40] It is perhaps illuminating at this point to compare this analysis of perfect competition and Pareto optimality in the two-parameter model with the corresponding analysis in the "states-of-the-world" model of Arrow [1] and Debreu [3]. It is well known that in the states model, a competitive equilibrium is Pareto optimal. But in the two-period version of this model, a competitive equilibrium would imply that there are at least two firms producing period 2 output for any given state. Thus there is no component in the output of a firm that is independent of the outputs of all other firms. If there were such residual risks, this would be a violation of the conditions required for perfect competition in the states model. But a monopoly in the states model is not necessarily a monopoly in the two-parameter model. In the states model, units of consumption in different states are not presumed to be perfect substitutes; but in the two-parameter model, all units of residual risk are perfect substitutes even though, since their realizations are independent of one another, they refer to outputs in different states of the world.

[41] In [2] and [13], respectively.

However, when the capital market is perfectly competitive, the production decisions of value maximizing firms are not Pareto optimal if individual firms in the risky industry have components of period 2 output that are independent of the outputs of all other risky firms. With these independent period 2 outputs, there are natural diversifying effects or external economies of scale—essentially the law of large numbers at work—that are not taken into account by perfectly competitive firms. As a result, aggregate output of such residual risks is less than would be required for a Pareto optimum.

It is well to conclude by noting that, at least in the two-parameter framework, the uncertainty analysis of competitive equilibrium is essentially the same as the certainty analysis. Specifically, in both cases, the capital market is perfectly competitive when conditions (C1)-(C6) hold. That is, the absence of perfect competition is not inherent in the existence of uncertainty, but rather, as in the certainty model, must arise from something like factor immobility, monopolistic access by individual firms to production techniques, or lack of information concerning the returns earned by given factors in different uses. Likewise, when a competitive equilibrium in the two-parameter model is not Pareto optimal, it is because of external economies of scale that are not taken into account by individual firms. But it is well known that such externalities cause similar problems in the perfect certainty model.

References

1. ARROW, K. J. "The Role of Securities in the Optimal Allocation of Risk-Bearing." *Review of Economic Studies*, Vol. 31, No. 2 (April 1964), pp. 91–96.
2. BLUME, M "Portfolio Theory: A Step Toward Its Practical Application." *Journal of Business*, Vol. 43, No. 2 (April 1970), pp. 152–73.
3. DEBREU, G. *The Theory of Value*. New York: John Wiley & Sons, Inc , 1959.
4. DUSAK, K. "Futures Trading and Investor Returns: An Investigation of Commodity Market Risk Premiums." Unpublished Ph.D. dissertation. University of Chicago, 1971.
5. FAMA, E. F. "The Behavior of Stock Market Prices." *Journal of Business*, Vol. 38, No. 1 (January 1965), pp. 34–105
6. ———. "Portfolio Analysis in a Stable Paretian Market." *Management Science*, Vol. 12, No. 5 (January 1965), pp. 404–19.
7. ———. "Risk, Return, and Equilibrium." *Journal of Political Economy*, Vol. 79, No. 1 (January/February 1971), pp. 30–55.
8. ———. "Risk, Return, and Equilibrium: Some Clarifying Comments." *Journal of Finance*, Vol. 23, No. 1 (March 1968), pp. 29–40.

9. ———, and Laffer, A. B. "The Number of Firms and Competition." *The American Economic Review,* Vol. 62, No. 4 (September, 1972).

10. ———, and Miller, M. H *The Theory of Finance.* New York: Holt, Rinehart and Winston, 1972.

11. ———, Fisher, L., Jensen, M., and Roll, R "The Adjustment of Stock Prices to New Information." *International Economic Review,* Vol. 10, No. 1 (February 1969), pp. 1–21.

12 Jensen, M. C. and Long, J. B., Jr. "Corporate Investment Under Uncertainty and Pareto Optimality in the Capital Markets." *The Bell Journal of Economics and Management Science,* Vol. 3, No. 1 (Spring 1972), pp. 151–74.

13 King, B F. "Market and Industry Factors in Stock Price Behavior." *Journal of Business,* Vol. 39, No 1 (Special Supplement, January 1966), pp 139–90

14. Lintner, J. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *The Review of Economics and Statistics,* Vol. 47, No. 1 (February 1965), pp. 13–37

15. Markowitz, H. *Portfolio Selection: Efficient Diversification of Investments* New York: John Wiley & Sons, Inc , 1959.

16. Mossin, J. "Equilibrium in a Capital Asset Market." *Econometrica,* Vol. 34, No. 4 (October 1966), pp. 768–83.

17. ———. "Security Pricing and Investment Criteria in Competitive Markets." *The American Economic Review,* Vol. 59, No. 5 (December 1969), pp. 749–56.

18. Press, S. J "Multivariate Stable Distributions." Mimeographed. University of Chicago, 1970.

19. Roll, R. *The Behavior of Interest Rates.* New York: Basic Books, 1970.

20 Tobin, J "Liquidity Preference as Behavior Toward Risk." *Review of Economic Studies,* Vol. 25, No. 1 (February 1958), pp. 65–85

21. Sharpe, W. F "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk." *Journal of Finance,* Vol. 19, No. 4 (September 1964), pp. 425–42.

22. ———. "A Simplified Model for Portfolio Analysis." *Management Science,* Vol. 9, No. 5 (January 1963), pp. 277–93.

23. Stiglitz, J. "On the Optimality of the Stock Market Allocation of Investment." *Quarterly Journal of Economics,* Vol 86, No 342 (February 1972), pp. 25–60

Nonhomogeneous production functions and applications to telecommunications

H. D. Vinod
Applied Statistics Department
Bell Telephone Laboratories

A form of nonhomogeneous production function is utilized to compute marginal productivities, various elasticities, optimum input ratios, and the like, for different levels of inputs and outputs. Such comparisons are relevant for labor negotiations, capital investment, and control by either a parent corporation or a government regulatory agency. This form of production function can be fitted by simple regression and allows variable returns to scale and variable elasticities of substitution.

The author illustrates applications of the function with two preliminary studies of Bell System data. The first study includes estimation of certain cost elasticities for the manufacture of sealed contacts by the Western Electric Co. The second deals with an aggregative production function for the Bell System, where it is found that the empirical evidence does not support the assertion that the Bell System is over-capitalized.

1. Introduction

■ In this paper we suggest a new formulation of the production function useful for making policy inferences from empirical studies.[1] Empirical computations of marginal productivities, various elasticities, optimum capital-labor ratios, isoquants, and the like, for different levels of inputs and outputs can be usefully compared, provided the functional form chosen does not assume away their differences.

We utilize a special case of a log-quadratic production function suggested by Kmenta[2] in a different context. It is a simple generalization of the Cobb-Douglas (C-D) formulation;[3] we simply add a product of logs of the two inputs (capital and labor). Such a function is linear in its parameters, which can be estimated by ordinary least squares. Recently, Griliches and Sargan[4] have also utilized special cases of the log-quadratic form.

Hrishikesh D Vinod received his B. Com (1959) from the Brihan Maharashtra College of Commerce, University of Poona, India, his M.A. (1961) from the Delhi School of Economics, and his Ph.D. (1965) from Harvard University He is currently a research econometrician in the Applied Statistics Department at Bell Laboratories, where he is engaged in productivity analysis and telephone usage forecasting.

The author wishes to thank R. L. Chaddha, J. A. Tischendorf, and R. B. Murphy of Bell Telephone Laboratories and E. F. Sudit and M. Burnside of the American Telephone and Telegraph Co for comments on an earlier version.

[1] The basic idea of nonhomogeneous production functions was presented by the author in 1968 (see [22]) The production function formulated here was developed for application by different Bell Telephone companies.

[2] See [14].

[3] See [6].

[4] In [9] and [19], respectively.

Section 2 discusses the mathematical properties of the proposed nonhomogeneous production function. In Subsection 2(1) we derive expressions for the marginal productivities, marginal elasticities, and marginal rate of substitution. We note that for small values of inputs the function may not be defined. Subsection 2(2) deals with the elasticity of substitution. The suggested form of the production function has variable returns to scale and variable elasticity of substitution (VES).

Section 3 discusses the empirical estimation. Subsection 3(1) discusses the computation of the optimum capital-labor ratio from empirical data. Subsection 3(2) deals with plotting the isoquants of an empirical production function. Subsection 3(3) deals with the new tool for economic policy.

Section 4 demonstrates applications of the methods detailed in Sections 2 and 3. An example dealing with certain Western Electric Co. data is given in Subsection 4(1), while an aggregative production function for the Bell System is discussed in Subsection 4(2).

2. The proposed nonhomogenous production function

■ Nonhomogeneous production functions are not new. For example, a homothetic function in Clemhout's paper and the VES functions in Revankar and Zellner-Revankar's work[5] are nonhomogeneous. The following form is convenient and somewhat uncommon. Let y denote output. Let $x_i > 0$ be the inputs such that x_1 = capital, x_2 = labor, and let a_k be the parameters. From now on we use subscripts i, j, and k with the convention that both i and $j = 1,2$, $i \neq j$, and $k = 0, \ldots, 3$ unless stated otherwise.

Consider a nonhomogeneous production function,

$$y = e^{a_0} x_1^{a_1 + a_3 \ln x_2} x_2^{a_2}, \tag{1}$$

which can be written as

$$\ln y = a_0 + a_1 \ln x_1 + a_2 \ln x_2 + a_3 \ln x_1 \ln x_2. \tag{2}$$

For brevity we shall not discuss the extension of (2) for three or more inputs. The example in Subsection 4(1) dealing with Western Electric Co. data uses the three-input version. When $a_3 = 0$, (1) becomes the well-known Cobb-Douglas function.[6] If the terms involving $(\ln x_1)^2$ and $(\ln x_2)^2$ are included, it is called the log-quadratic formulation, first used by Kmenta in a different context.[7]

☐ **(1) Marginal elasticities, productivities, etc.** The marginal elasticity of x_i is denoted by ϵ_i and defined to be $(\partial \ln y / \partial \ln x_i)$:[8]

$$\epsilon_i = a_i + a_3 \ln x_j. \tag{3}$$

---

[5] [5], [18], and [25], respectively.

[6] Equation (2) is a special case of the log-quadratic form. When the data are collinear, it is impossible to include so many terms in the regression equation The present author had attempted to fit many functional forms that are linear in parameters to the examples reported in this paper. However, these were rejected on statistical grounds because of multicollinearity, serial correlation [24], etc. Some of these forms are listed in note 9 below.

[7] See [14].

[8] Frisch was the first to define the concept of marginal elasticity (see [8], p. 77). Its use for nonhomogeneous functions was determined by the author in [22]. The

The scale elasticity is $\epsilon = \epsilon_1 + \epsilon_2$, hence

$$\epsilon = a_1 + a_2 + a_3 \ln(x_1 x_2), \tag{4}$$

which measures the returns to scale behavior. The marginal productivities are as follows:

$$f_1 = \mathrm{MP}_1 = \frac{y}{x_1}\epsilon_1, \tag{5}$$

$$f_2 = \mathrm{MP}_2 = \frac{y}{x_2}\epsilon_2. \tag{6}$$

The marginal rate of substitution is

$$\mathrm{MRS}_{12} = x_2\epsilon_1/x_1\epsilon_2. \tag{7}$$

Assuming $y > 0$ and $x_i > 0$, $(i = 1,2)$, the sign of the marginal productivities $f_i (i = 1,2)$ can be negative if and only if ϵ_i is negative.[9] The condition for $\epsilon_i > 0$ is that x_j should satisfy

$$-a_i < a_3 \ln x_j. \tag{8}$$

Note that (8) defines a region of variation of x_j.[10]

An isoquant is defined as the locus of all combinations of x_1 and x_2 which yield a specified output level. The condition for convexity (to the origin) of the isoquants is[11]

$$(f_2^2 f_{11} - 2f_1 f_2 f_{12} + f_1^2 f_{22}) < 0, \tag{9}$$

where $f_{ij} = (\partial^2 y/\partial x_i \partial x_j)$. Now, (9) reduces to $(y^3/x_1^2 x_2^2)$ $[\epsilon_1\epsilon_2(\epsilon_1 + \epsilon_2 + 2a_3)] > 0$. Hence given that y, x_i, and a_3 are positive, a sufficient condition for the convexity of isoquants is $\epsilon_i > 0$. However, all isoquants need not be convex everywhere and negative ϵ_i can help us locate lack of convexity. This may be

---

theorem that the scale elasticity is a sum of all marginal elasticities is proved by Frisch. It is an undesirable property of (2) that when $a_3 > 0$, ϵ_i increases indefinitely with the other input. The marginal productivities $f_i = (y/x_i)\epsilon_i$ for (2) do decrease with x_i. If we want ϵ_i to decrease with x_i, the following formulations may be appropriate (let capital letters denote the logs):

(i) $Y = a_0 + a_1[X_1 - (1/X_1)] + a_2[X_2 - (1/X_2)] + a_3 X_1 X_2$,

(ii) $Y = a_0 + a_1 X_1 + a_2 X_2 + a_3[X_1 - (1/X_1) - (1/X_2)]$,

(iii) $Y = a_0 + a_1(X_1 + \ln X_1) + a_2(X_2 + \ln X_2) + a_3 X_1 X_2$, and

(iv) $Y = a_0 + a_1 X_1 + a_2 X_2 + a_3[X_1 X_2 + \ln X_1 + \ln X_2]$.

None of these functions were statistically acceptable for the data reported in the examples given. In the absence of experimental data, it is difficult to choose among them on economic considerations alone. None of them should be used for very wide ranges in inputs and outputs.

[9] Differentiating (5) and (6) with respect to x_1 and x_2 respectively, we find that the marginal productivity f_i is increasing, decreasing, or constant as long as ϵ_i^2 is greater than, less than, or equal to ϵ_i $(i = 1,2)$ When we assume $\epsilon_i > 0$ we have further simplification: f_i is increasing, decreasing, or constant as long as $\epsilon_i > 1$, $\epsilon_i < 1$ or $\epsilon_i = 1$.

[10] Given a_3 as positive (negative) and a_i as negative (positive) as long as x_j is greater (smaller) than $\exp(-a_i/a_3)$, the marginal productivities and marginal elasticities will be positive. The case where both a_i and a_3 are positive satisfies (8) for any $\ln x_j$, while the case where both a_i and a_3 are negative cannot satisfy (8) for any $\ln x_j$, $(x_j \geq 1)$.

[11] See Allen [1], p. 338.

important because the optimum capital-labor ratio obtained in the usual manner may yield maximum rather than minimum costs when isoquants are concave, as we shall see in Subsection 3(1).

☐ **(2) Elasticity of substitution.** From the definition of the elasticity of substitution[12] it is easy to show that

$$\sigma = \frac{\epsilon_1 + \epsilon_2}{\epsilon_1 + \epsilon_2 + 2a_3} = \frac{a_1 + a_2 + a_3 \ln(x_1 x_2)}{a_1 + a_2 + a_3[\ln(x_1 x_2) + 2]}. \quad (10)$$

This cannot be directly obtained from a similar formula for log-quadratic functions given by Sargan,[13] because his formula assumes $\ln x_1 = \ln x_2 = 0$; it is intended to be evaluated at the arithmetic mean of the sample values of x_1 and x_2 after rescaling. Equation (10) is true for any x_1 and x_2 in the sample.

Thus we have variable elasticity of substitution because ϵ_1 and ϵ_2 are variable with the levels of x_1 and x_2. Also note that $\epsilon_1 + \epsilon_2 \gtrless 0$ and $a_3 \gtrless 0$ imply $\sigma < 1$. Other empirical studies[14] lead us to expect that most production functions based on time series data have $\sigma < 1$. The sign of a_3 may give a simple check on the appropriateness of the corresponding function.

Since ϵ_1, ϵ_2, and σ are elasticities, it is obvious that they are free from units of measurement. Other nonhomogeneous formulations can be considered along similar lines. Our nonhomogeneous model has interesting implications for estimation of the elasticity of total cost, average cost, marginal cost, supply, and so on. These are discussed more conveniently in Subsection 4(1) in the context of an application to Western Electric Co. data.

3. Empirical estimation of nonhomogenous production functions

■ Let the empirical data consist of a time series or a cross section of n observations on y_t, x_{1t}, and x_{2t} $(t = 1, \ldots, n)$ variables. We can introduce, for convenience, an additive error term u_t in (2), assuming that its mean and variance are $E(u) = 0$, $\mathrm{Var}(u) = \sigma_u{}^2 I$, where I is the identity matrix. The formulation (2) is linear in its parameters a_k $(k = 0, \ldots, 3)$, and therefore ordinary least squares can be used to obtain estimates of the a_k, denoted by $\hat{a}_k$.

The Student's t-statistic for $\hat{a}_3$ can be used to test whether $\hat{a}_3$ is significantly different from zero. If at a certain level (say 5 percent) $\hat{a}_3$ is not significantly different from zero, we may reject the nonhomogeneous formulation (2), and use, for example, the C-D function. Equation (10) reveals that when $a_3 \neq 0$, σ will be constant if and only if the product $(x_{1t}x_{2t})$ is constant for all $t = 1, \ldots, n$. Thus the Student's t-test on a_3 will suggest that the constant elas-

---

[12] The concept of elasticity of substitution was first discussed by Hicks [10], who also mentioned the possibility that a production function may have constant or variable elasticity of substitution. An equivalent definition to (10) is

$$\sigma = \left(\frac{1}{x_1 f_1} + \frac{1}{f_2 x_2}\right) \Big/ -\left(\frac{f_{11}}{f_1^2} - 2\frac{f_{12}}{f_1 f_2} + \frac{f_{22}}{f_2^2}\right),$$

where f_{ij} is $(\partial^2 f/\partial x_i \partial x_j)$ and $i,j = 1,2$

[13] In [19].

[14] See Arrow et al [2] or Lovell [16].

ticity of substitution (CES) function[15] is inappropriate whenever $\hat{a}_3$ is significantly different from zero. This conclusion may not always follow, because the CES function is not a special case of our nonhomogeneous function.[16]

Note that, using the regression estimates $\hat{a}_k$ and the formulas in the previous section, one can derive empirical estimates of MP_i, ϵ_i, ϵ, MRS_{ij}, and σ for each observed level of y_t, x_{1t}, and x_{2t} $(t = 1, \ldots, n)$. These are illustrated in Section 4.

☐ **(1) Computation of the optimum capital-labor ratio.** To find the optimum capital-labor ratio, denoted by $\mathrm{opt}(x_1/x_2)$, we need additional data on unit rentals (interest rate) on capital, denoted by p_{1t}, and unit wages, denoted by p_{2t} $(t = 1, \ldots, n)$. The first-order conditions for the optimum are that the marginal rate of substitution equals the price ratio, also stated as

$$(f_1/p_1) = (f_2/p_2). \tag{11}$$

Substitution of (5) and (6) into (11) gives the optimum capital-labor ratio by

$$\mathrm{opt}(x_1/x_2) = p_2\epsilon_1/p_1\epsilon_2. \tag{12}$$

The second-order conditions for cost minimizing optimum are given by (9), and negative ϵ_1 or ϵ_2 may lead to their failure. The second-order conditions for profit maximizing optimum are more stringent: $p\,f_{11} < 0$ and $p\,f_{22} < 0$, where p denotes the market price of y. This would require diminishing returns and/or market imperfections. It would be convenient to interpret (12) as a formula for obtaining the cost minimizing optimum.

The behavior of the empirical $\mathrm{opt}(x_1/x_2)$ involving estimates $\hat{a}_k$ can be traced for all observations $t = 1, \ldots, n$ and compared to the actual capital-labor (x_1/x_2) ratio.

☐ **(2) Plotting the isoquants.** To visualize our nonhomogeneous production function (2) it may be helpful to plot the isoquants, i.e., the locus of (x_1, x_2) for certain fixed values of y, denoted y_*. Rewriting (2) we obtain the equation of the isoquants

$$x_i = \exp[(\ln y_* - a_0 - a_j \ln x_j)/(a_i + a_3 \ln x_j)]. \tag{13}$$

These isoquants are downward sloping provided $(dx_i/dx_j < 0)$, i.e.,

$$a_1 a_2 + a_3(\ln y_* - a_0) > 0. \tag{14}$$

---

[15] In [2].

[16] The substitution behavior between the inputs in economics cannot be directly observed by experimentally varying the inputs, keeping the output fixed. Hence a functional form that allows flexible substitution behavior exemplified by VES between the inputs, such as (1) or (2), seems to have less specification bias than the more rigid forms, such as C-D or CES. For some of the difficulties of economic interpretation associated with the CES production functions, see Dobell [7] and other articles in the same issue A detailed justification and additional references for VES formulation may be found in Revankar [18]. Lovell [16] had also recommended VES formulations, while retaining the assumptions of homogeneity and neutrality of technological change. If nonlinear regression methods are not available, CES functions cannot be fitted unless we assume that the marginal productivity theory holds in the capital and/or labor markets. Our function (2) does not require nonlinear regression.

This condition is equivalent to ϵ_1 and $\epsilon_2 > 0$ derived from (2); (3), and (9). Substituting $\hat{a}_k$ in (14) we note that the condition holds for y_* values greater (smaller) than $\exp[\hat{a}_0 - (\hat{a}_1\hat{a}_2/\hat{a}_3)]$ provided $\hat{a}_3$ is positive (negative). If such y_* values are unrealistic we may suspect the estimates $\hat{a}_k$. Otherwise we proceed to plot the empirical isoquants.

We are usually interested in a small range of values of y, x_1, and x_2. This range may be slightly wider than the corresponding observed range. In drawing the isoquants from $\hat{a}_k$ in (13) it helps to restrict one's numerical evaluations to the above-mentioned range.[17]

☐ **(3) A tool for descriptive and policy implications.** The empirical work is not completed when an appropriate production function is fitted. Our main suggestion is that we can further compute for each of the data points ($t = 1, \ldots, n$) empirical estimates of certain economically meaningful constructs—MP_i, MRS_{ij}, ϵ_i, ϵ, σ, the isoquants, $\mathrm{opt}(x_1/x_2)$, etc.—provided the functional form chosen does not mathematically preclude the variation in these constructs. Since ϵ_i is constant for the C-D function, σ is constant for the CES, and the homothetic functions have parallel isoquants, we prefer our nonhomogeneous formulation (2). Further, we can plot, compare, rank-order, and classify all the data points with respect to the constructs listed above. This rich source of information appears to have been ignored in the literature, and is suggested here as a tool for (1) better description of economic reality and (2) policy recommendations. A better understanding of economic reality is essential for meaningful policy-making; hence we refer to our suggestion briefly as a new tool for policy. Since a policy tool should be simple to use, we have avoided formulations that may require nonlinear regression. Some examples of the use of this tool are mentioned below. In the work of Sargan and Griliches-Ringstad[18] the marginal productivities and the elasticity of substitution are estimated at the sample mean points of the input variables, but there is no comparison of the elements within the sample.

In labor negotiations and testimonies before the Federal Pay Board, the term "productivity" of labor is often vaguely used. The marginal productivity and elasticity of labor (MP_2, ϵ_2) are relevant concepts. Our tool makes it possible to empirically estimate, compare, and forecast these and other constructs from appropriate time-series data.

Another example of policy inference is a circumstance in which a parent corporation or a government regulatory agency needs to evaluate and control the performance of certain other corporations. A cross-sectional study of these corporations can provide a ranking of each corporation with respect to several economically meaningful quantities listed above, thereby providing a tool for judging their performance.

---

[17] The author has written a computer program which requires specification of the ranges of x_1 and x_2 values. It then evaluates x_2 for a given set of x_1 values, and y_* by (13). If such x_2 becomes too large (small) and falls outside the range, the program sets x_2 equal to the largest (smallest) admissible value of x_2. There are computer programs which can directly plot arbitrary surfaces. The sampling variation in $\hat{a}_k$ can be used to construct a confidence band within which each isoquant would lie 95 percent of the time. See note 19 for further details. The standard errors of $\hat{a}_k$ can be readily estimated.

[18] [19] and [9], respectively.

A comparative study of empirical MP_i, ϵ, etc., may make possible a review of past policy decisions, evaluation of plans for future policy-making, and evaluation of forecasts. Additional capital may or may not be invested in a particular year or region, depending on MP_i, ϵ_i values. More examples of the use of the proposed tool can be conceived. However, for want of less aggregative, more reliable and possibly confidential data, it is not possible to give a more convincing practical illustration. Before making strong policy recommendations we also should remember that the regression coefficients $\hat{a}_k$ used in our computations are subject to random variation.[19]

The empirical examples of the following section illustrate the calculations involved in the use of our tool, especially for a better *description* of economic reality. Some speculative hints for policy are included.

4. Examples of empirical studies

■ In this section we illustrate with examples some of the methods developed in earlier sections. The first subsection uses manufacturing data from the Western Electric Co. to deal with the estimation of cost- and supply-related elasticities. The second subsection discusses our estimation of a nonhomogeneous production function for aggregative data of the Bell System.

☐ **(1) Microeconomic study of Western Electric Co.'s manufacture of sealed contacts.** Thousands of units of an electronic device called a sealed contact are produced by the Western Electric Co. There are data available for the monthly time series of output y and three inputs: x_1 = capital, x_2 = labor, and x_3 = engineering, on the basis of 59 observations. The input data were carefully selected to refer to the specific output. A nonhomogeneous production function of the following form is proposed:

$$\ln y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_1X_2 + a_5X_1X_3 + a_6X_2X_3, \quad (15)$$

where $X_i = \ln x_i$ $(i = 1, \ldots, 3)$.

From our data, the values of $\hat{a}_i$ $(i = 0, \ldots, 6)$ and the corresponding absolute values of the Student's t-statistic (in parentheses) are, respectively: -11.577 (7.29), -4.918 (2.56), 1.999 (6.27), 11.820 (3.56), 0.976 (2.60), 0.583 (5.97), and -1.930 (2.96). Thus at the 5-percent level of significance all regression coefficients are statistically significant.

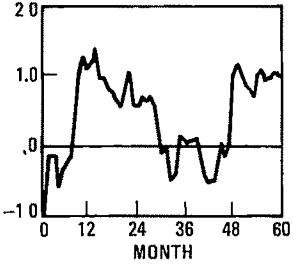
---

[19] The formulas for MRS_{12}, opt (x_1/x_2), or the isoquant (13) involve $\hat{a}_k$ values in both the numerator and the denominator. This leads to a ratio of correlated random variables which may be assumed to be normally distributed. The ideas given in Hinkley [11] may be used, or more simply one can use an approximation (see Barlett [3], p. 256).

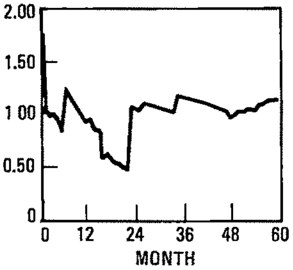
FIGURE 1

MARGINAL ELASTICITIES

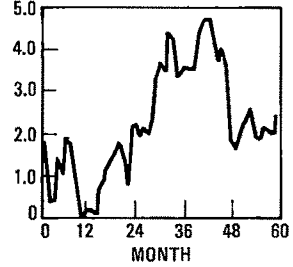
(a) MARGINAL ELASTICITY OF CAPITAL (ϵ_1)



(b) MARGINAL ELASTICITY OF LABOR (ϵ_2)



(c) MARGINAL ELASTICITY OF ENGINEERING (ϵ_3)



Now we turn to two kinds of policy implications which can be derived from this study.[20]

Employment policy

Capital input is mainly represented by machinery that produces the sealed contacts. Figure 1(a) reflects the ability of a graph of the marginal elasticity of capital to pinpoint the months when the machines were relatively idle (months 32, 33) and the months when the machines were used heavily (months 12, 49, etc.). This enhances our confidence in the numerical estimates of marginal productivity and marginal elasticity of capital to determine the appropriate level of capital employment. Similar comments can be made regarding the labor input.

The marginal elasticity of the engineering input is given in Figure 1(c). From Figure 1 note that a 1-percent increase in the employment of engineering personnel has a relatively high proportionate yield. Our analysis suggests that only a mild increase should occur in the employment of this input, after its price (not indicated here) is taken into account.

Pricing policy

For the industrial engineers of the Western Electric Co. the price of the sealed contact is one of the important policy variables. Further, the amount of the quantity discount offered to the customers needs to be determined. The price policy may have an impact on the quantities demanded and produced via the demand curve. Derivation of the empirical demand curve is well known and outside the scope of this paper. An empirical study of the response of total cost (C), marginal cost (MC), and average cost (AC) to the varying quantities of outputs produced is appropriate, since these variables are relevant to the pricing decision.

Before we turn to the data, we should complete a short review of certain known results of the theory of production. The reader may refer to Frisch[21] for a fuller discussion. Our aim is to show that an interesting relationship exists between certain cost elasticities and the scale elasticity (ϵ). Assuming fixed prices of the three inputs and assuming no monopsony in the market for the inputs, it is well known that the first-order conditions for the minimization of the total costs are

$$p_i = (\partial C/\partial y) f_i, (i = 1,2,3), \qquad (16)$$

where p_i is the price of ith input, C total cost, and f_i the marginal productivity of ith input. Substitution of $f_i = (y/x_i)\epsilon_i$ and $C = p_1x_1 + p_2x_2 + p_3x_3$ in (16) yields

---

[20] The relevance of production functions to the managers of manufacturing plants and to the industrial engineers for cost estimation, cost prediction, capacity planning, choice of the manufacturing source, etc., is discussed by Lele and O'Leary [15] of the Western Electric Co. They found certain difficulties in using conventional production functions from economic theory. They had discussed the difficulties with me in mid-1971, and this section grew partly out of my desire to give an economist's reply to their questions.

[21] See [8], pp. 167–69.

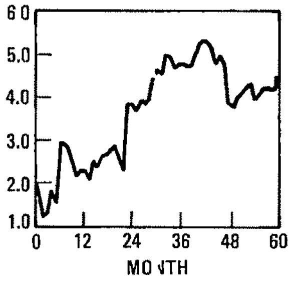
$$El(\mathrm{C}) = \frac{y}{\mathrm{C}}\frac{\partial \mathrm{C}}{\partial y} = \frac{1}{\epsilon_1 + \epsilon_2 + \epsilon_3} = \frac{1}{\epsilon}, \tag{17}$$

where *El* denotes elasticity.

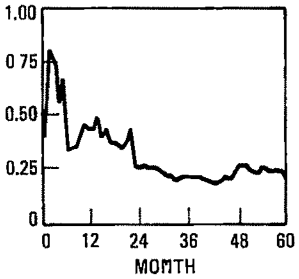
Thus the elasticity of the total cost with respect to output is simply the reciprocal of the scale elasticity ϵ (Figure 2). Data on average cost are appropriate here. The numerical evaluation of this reciprocal elasticity, given in Figure 2(b) should be interpreted with an understanding that the so called "total costs" do not include materials, taxes, distribution, or overhead costs. Thus, from Figure 2(b), a 1-percent increase in the quantity produced during the 3rd month leads to a 0.8-percent increase in total costs. This elasticity reduces to 0.25 during the 22nd month, and remains around that level till the end of the time period covered. This indicates significant cost reductions during the 3rd to 22nd month.

FIGURE 2

(a) SCALE ELASTICITY



(b) RECIPROCAL OF SCALE ELASTICITY



Certain other elasticities can also be estimated from the empirical scale elasticity after mathematical manipulations of (17), as in Frisch. We have

$$El(\mathrm{AC}) = (1/\epsilon) - 1, \tag{18}$$

$$El(\mathrm{MC}) = El(\mathrm{AC}) - El(\epsilon), \tag{19}$$

and

$$El(\text{supply with respect to price}) = 1/El(\mathrm{MC}). \tag{20}$$

The elasticities in (18) and (19) are with respect to the quantity produced. The $El(\epsilon)$ is difficult to express analytically for our functional form. However, if we do not need the estimate of these values at the end points of the data, we can calculate it by numerical methods (e.g., by IBM's scientific subroutine called DDGT3).[22] These numerical estimates change sign several times to correspond with the ups and downs in ϵ, and are judged to be not very useful for this application. We have included the expressions (19) and (20) in this paper, because they may be useful in other examples.

The exploratory example of this section shows that a great deal of the neoclassical theory of production can be subjected to empirical estimation, although such a possibility is not emphasized in the literature. Nonhomogeneous production functions seem to be particularly suitable for such empirical work.

☐ **(2) Aggregate production function for the Bell System.** An exploratory study of the Bell System data for the time period 1947–1970 was made using the ideas discussed earlier in this paper. Let y = net value added output series appropriately deflated to exclude the effects of inflation, x_1 = the net capital stock employed, and x_2 = labor input in man-hours.[23] The fitted production function is

$$\begin{aligned} \ln y = {} & \underset{(8.501)}{69.8541} - \underset{(7.896)}{6.40795} \ln x_1 \\ & - \underset{(8.671)}{10.5900} \ln x_2 + \underset{(9.131)}{1.09574} (\ln x_1 \ln x_2), \end{aligned}$$

---

[22] See [12].

[23] We have not given additional details regarding the data partly because they are private and partly because our work in refining the data base is somewhat incomplete.

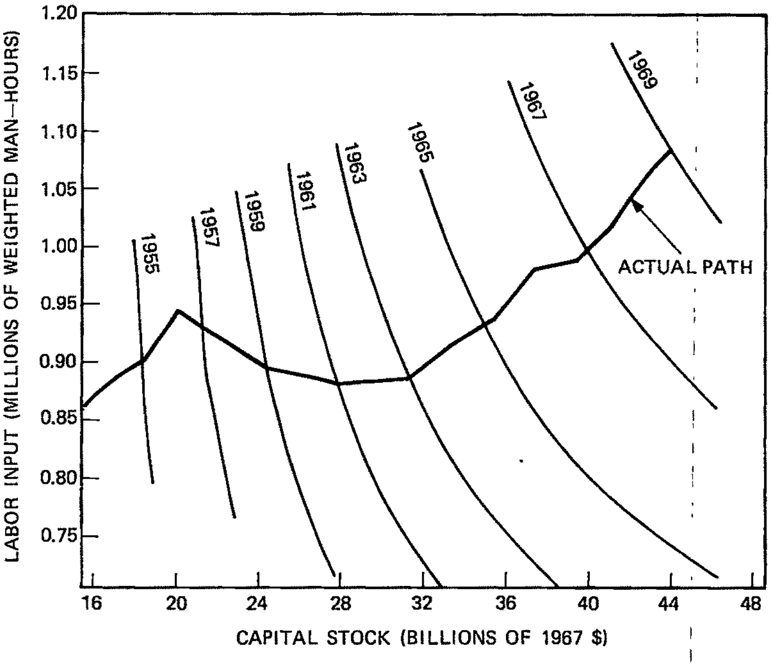
TABLE 1

MARGINAL PRODUCTS AND ELASTICITY MEASURES, 1947–1970
(NONHOMOGENEOUS PRODUCTION FUNCTION–NET VALUE–ADDED VERSION)

	MARGINAL PRODUCT OF CAPITAL	MARGINAL PRODUCT OF LABOR	MARGINAL ELASTICITY OF CAPITAL	MARGINAL ELASTICITY OF LABOR	SCALE ELASTICITY	ELASTICITY OF SUBSTITUTION
	f_1	f_2	ϵ_1	ϵ_2	ϵ	σ
	(1)	(2)	(3)	(4)	(5)	(6)
1947	0.158265	−1.451450	0.802871	−0.563984	0.238888	0.098293
1948	0.153702	−1.086850	0.835654	−0.402418	0.433236	0.165060
1949	0.147941	−0.806541	0.865190	−0.300763	0.564427	0.204806
1950	0.154764	−0.676782	0.877466	−0.232481	0.644985	0.227391
1951	0.175840	−0.619529	0.920884	−0.198225	0.722660	0.247984
1952	0.181338	−0.300996	0.971151	−0.093729	0.877422	0.285907
1953	0.179050	0.028189	1.008810	0.008705	1.017520	0.317082
1954	0.178084	0.273202	1.029393	0.082453	1.111850	0.336584
1955	0.184281	0.617585	1.051780	0.172948	1.224720	0.358504
1956	0.193461	0.981133	1.096960	0.262201	1.359160	0.382793
1957	0.190317	1.396260	1.076790	0.340372	1.417170	0.392714
1958	0.180637	1.867700	1.051900	0.424301	1.476200	0.402489
1959	0.182596	2.374370	1.035920	0.488454	1.524380	0.410236
1960	0.185058	2.853480	1.030870	0.544921	1.575800	0.418285
1961	0.182852	3.404040	1.022330	0.610028	1.632350	0.426889
1962	0.183098	4.056830	1.023940	0.681975	1.705920	0.437707
1963	0.184232	4.681310	1.027290	0.743853	1.771140	0.446962
1964	0.193999	5.381670	1.059260	0.810945	1.870200	0.460450
1965	0.200491	6.182420	1.095220	0.893355	1.988570	0.475729
1966	0.226332	7.107690	1.140360	0.943219	2.083580	0.487380
1967	0.236381	8.198900	1.150030	1.003340	2.153370	0.495614
1968	0.255848	9.092000	1.179300	1.043610	2.222910	0.503560
1969	0.285140	10.184600	1.242440	1.105910	2.348350	0.517277
1970	0.293575	11.221800	1.282490	1.182080	2.464570	0.529327

FIGURE 3

ISOQUANTS FOR THE BELL SYSTEM NONHOMOGENEOUS FUNCTION
(NET VALUE–ADDED VERSION)



where the numbers in the parentheses indicate the absolute values of the Student's t-statistic. The multiple correlation coefficient $R^2 = 0.9966$; the Durbin-Watson statistic $= 1.20$; the value of the first-order serial correlation among estimated residuals is 0.36.

The numerical results are given in Table 1, and a plot of the isoquants appears in Figure 3. In the early postwar years certain nonconvexities occur in the isoquant map until the year 1955. The Bell System began the postwar period with a highly labor-intensive structure of production illustrated by the fact that few manual telephone central offices were established to keep up with the postwar "pent-up" demand. The shortage of the telephone equipment was serious and it took many years of very strong growth in capital investment to make up for it. The situation was remedied by the year 1955; however, the isoquant for that year in Figure 3 is almost vertical. The marginal productivity and marginal elasticity of labor in the early postwar years are negative in Table 1 for the same reason. This may be due to data aggregation, or inappropriateness of our basic model. We view these negative values as providing us with additional historical information about the postwar years which we may otherwise fail to notice.

The marginal elasticity of labor in Table 1 has remained lower than that of capital throughout the time period covered; however, the values of the two variables seem to be coming closer to each other as the time passes. Interpretation of the marginal productivities is similar to the previous example and is omitted.

The scale elasticities (ϵ) given in Table 1 include not only the effects of largeness of the scale of operations, but also the effects of technological change.[24] The separation of ϵ into two separate components measuring scale and technological change is difficult and may not be possible, because of a close tie between technology, size, and research and development expenditures in the telephone business; the size and growth of the telephone business seems to be intimately related to the speed and extent of technological change. Also, a great deal of technology has been internally generated; some of it is embedded in the inputs.

Econometric estimation of neutral technological change is often based on Solow's approach,[25] where time is used as a surrogate for technology. There are several visible instances of labor-saving technological change in the Bell System, such as Direct Distance Dialing (DDD); hence a model that assumes neutral technological change may not be appropriate. We would prefer to use hard engineering data if we must separate the effects of size and technology. A statistical separation when most of our variables are highly correlated with each other can be misleading. The collection of the engineering data seems to be a major undertaking.

The 1970 value of ϵ in Table 1 indicates that a 1-percent increase in both labor and capital inputs would lead to an increase in output of about 2.5 percent. Since the scale elasticity is a ratio of percentage changes of output to inputs, it is similar to the "total factor productivity" measure of growth, which is defined by Kendrick[26] as a change in the ratio of output to inputs. From Table 1, ϵ increased from about 1.22 in 1955 to about 2.46 in 1970, which may indicate a doubling of overall productivity. Doubling in 15 years amounts to an approximate annual rate of growth of 4.75 percent per year.

---

[24] The expression "scale" elasticity is somewhat unfortunate, and I am tempted to name it "tech-scale" elasticity to emphasize that both technology and scale are included.

[25] See [21].

[26] In [13].

The elasticity of substitution (σ) given in Table 1 reflects the ease with which capital is substituted for labor in response to a change in the relative productivity of the factors. The results indicate increasing substitutability over time. The fact that the elasticity of substitution is always lower than unity suggests that the Cobb-Douglas (C-D) formulation, which restricts it to unity, may not be adequately representative of the Bell System production process.

A computation of the optimal capital-labor ratio discussed in Subsection 3(1) is possible for the C-D model, and shows that the actual capital-labor ratio is much smaller than the optimum. These results are not given because of the possible inappropriateness of the C-D model. For the nonhomogeneous production function, the optimal capital-labor ratio is not defined for the period prior to 1955[27] for lack of convex isoquants. We have devised computations of capital cost indexes and labor wage indexes, but they are not available. Even so, we have written these numbers down for what they are worth as an illustration of the model. A comparison of the actual and the optimal ratios given in Table 2 for 1960–1970 suggests that there is empirical support for the statement that the Bell System is not overcapitalized. But this can not be taken as an "empirical result" because of the preliminary nature of the data and of the fitting of the equation to these data.

TABLE 2

OPTIMAL VS. ACTUAL CAPITAL–LABOR RATIOS (NONHOMOGENEOUS PRODUCTION FUNCTION)

YEAR	ACTUAL	OPTIMAL
1960	29.1703	50.9916
1961	31.1985	47.5220
1962	33.2667	44.9707
1963	35 0919	42.7670
1964	36.2350	44.9158
1965	37.8042	45.6654
1966	37.9675	49.4996
1967	39.7563	50.6892
1968	40.1574	50.8842
1969	40.1270	57 1108
1970	41.4716	59.4449

5. Concluding comment

■ In this paper we have attempted to show the power of nonhomogeneous production functions from the viewpoint of microeconomic, as well as macroeconomic, policy. It appears that a detailed analysis of each observation should follow the statistical fitting of production functions. The chosen form of nonhomogeneity and the examples are mainly illustrative. We have a tool for empirically testing the ancient concept of "gold plating" or "rate-base padding."

References

1 Allen, R. G. D. *Mathematical Analysis for Economists.* London: Macmillan & Co., 1956

2. Arrow, K. J., Chenery, H. B., Minhas, B. S., and Solow, R. M., "Capital-Labor Substitution and Economic Efficiency." *The Review of Economics and Statistics,* Vol. 43, No. 3 (August 1961), pp. 225–50.

3. Bartlett, M. S. *Stochastic Processes.* Cambridge: Cambridge Univ Press, 1960.

4. Bodewig, E. *Matrix Calculus.* New York: North-Holland Publishing Co., 1959

5. Clemhout, S. "The Class of Homothetic Isoquant Production Functions." *Review of Economic Studies,* Vol. 35, No. 1 (January 1968), pp. 91–104.

6. Cobb, C. W. and Douglas, H "A Theory of Production." *The American Economic Review,* Vol. 18, Supplement (March 1928), pp. 139–65.

7. Dobell, R. "A Symposium on CES Production Functions: Extensions and Comments, Introductory Comments." *The Review of Economics and Statistics,* Vol. 50, No. 4 (November 1968), pp. 443–45.

8. Frisch, R. *Theory of Production.* Chicago · Rand McNally & Co , 1965.

9. Griliches, Z and Ringstad, V. *Economics of Scale and the Form of the Production Function: An Econometric Study of Norwegian Manufacturing Establishment Data.* New York: North-Holland Publishing Co., 1971.

---

[27] See Subsection 3(1).

10. HICKS, J. R. *The Theory of Wages.* London: Macmillan & Co., 1932.
11. HINKLEY, D. V "On the Ratio of Two Correlated Normal Variables." *Biometrika,* Vol. 56, No 3 (December 1969), pp 635–39
12. International Business Machines Corp. "System/360, Scientific Subroutine Package" (360A-CM-03X) Version III, Publication No. H20-0166-5, White Plains, N Y, 1968.
13. KENDRICK, J W. *Productivity Trends in the United States* National Bureau of Economic Research Princeton Princeton Univ. Press, 1961
14. KMENTA, J. "On the Estimation of the CES Production Function." *International Economic Review,* Vol. 8, No. 2 (June 1967), pp 180–89
15. LELE, P. T. AND O'LEARY, J. W. "Application of Production Functions in Management Decisions." American Institute of Industrial Engineers, *AIIE Transactions,* Vol. 4, No. 1 (March 1972), pp. 36–42.
16. LOVELL, C. A. K. "Capacity Utilization and Production Function Estimation in Postwar American Manufacturing." *Quarterly Journal of Economics,* Vol. 82, No. 327 (May 1968), pp. 219–39.
17. MARSCHAK, J AND ANDREWS, W. H., JR. "Random Simultaneous Equations and the Theory of Production." *Econometrica,* Vol. 12, Nos 3–4 (July–October 1944), pp. 143–205.
18. REVANKAR, N. "A Class of Variable Elasticity of Substitution Production Functions" *Econometrica,* Vol. 39, No. 1 (January 1971), pp 61–71.
19. SARGAN, J D Part 5, in P. R. G. Layard et al., eds., *Qualified Manpower and Economic Planning,* Baltimore: Penguin Books, Inc., 1972.
20. SATO, R. AND BECKMANN, M. J. "Neutral Inventions and Production Functions." *Review of Economic Studies,* Vol. 35, No 1 (January 1968), pp. 57–66.
21. SOLOW R. M "Technological Change and the Aggregate Production Function." *The Review of Economics and Statistics,* Vol. 39, No. 3 (August 1957), pp 312–20.
22. VINOD, H. D. "Bounds on Returns to Scale for Arbitrary Nonhomogeneous Production Functions and an Application to the Leontief System" Paper presented at Econometric Society meeting, Evanston, Ill., December 1968.
23. ———. "Econometrics of Joint Production." *Econometrica,* Vol. 36, No. 2 (April 1968), pp. 322–36.
24. ———. "A Generalization of the Durbin-Watson Statistic for Higher Order Autoregressive Processes." Unpublished manuscript, 1971.
25. ZELLNER, A. AND REVANKAR, N. S. "Generalized Production Functions." *Review of Economic Studies,* Vol. 36, No 2 (April 1969), pp 241–50.

Valuation models in regulation

Blaine E. Davis

Corporate Planning Organization
American Telephone & Telegraph Company

and

F. T. Sparrow

Associate Professor of Economics and Operations Research
The Johns Hopkins University

This paper surveys and critiques the financial models of the firm used in rate of return regulation. The increasingly wide assortment of models used in rate of return cases might appear to promise a rich classification of contrasting assumptions and a diverse catalogue of model constructs. Nothing could be further from the findings of this paper. It is shown that all models considered have assumed either explicitly or implicitly the financial market conditions used in the development of the original Modigliani and Miller (M-M) model of 1961. Thus, the consequences, criticisms, and attributes of the Modigliani-Miller financial world seem to represent the current state of rate of return modeling.

1. Introduction

■ The objective of this paper is to survey and critique the financial models of the firm used in rate of return determinations for regulated public utilities. Such a study seems to be called for, since the degree of uncritical acceptance of such models by both sides of the proceedings frequently does not accurately mirror the diversity of opinion over the correctness of the constructs found in the financial literature.

The paper will first present a financial model of the firm that might be thought of as describing the "state of the art" in that it incorporates most of the recent thinking of financial analysts either explicitly in the formulation, or implicitly in the discussion of the model's characteristics. Next follows a section on the peculiarities of the regulated firm with suggestions as to how the model must be modified to reflect these facts of life. Finally, using the modified construct as a standard of comparison, the various financial models

---

Blaine E. Davis received the B S. degree in electrical engineering from the University of Maine at Orono (1964). From The Johns Hopkins University he was awarded the M.S degree in operations research and the Ph.D. degree in economics and operations research (1965 and 1969, respectively). Dr. Davis is Manager—Economic and Demand Studies in the A.T.&T. Company's Corporate Planning Organization. His current research activities are concerned with econometric modeling of the public utilities sector to analyze alternative regulatory policies.

F. T. Sparrow received the B.S. degree in geology from the University of Michigan (1953), the M B.A degree in managerial economics from Cornell University (1956), and the Ph.D. degree in economics and operations research from the University of Michigan (1962). He is currently Associate Professor of Economics and Operations Research at The Johns Hopkins University, where his research interests include the applications of capital theory to financial markets and the issues of efficiency and equity in the selection of optimal regulatory policies used in rate of return determinations are analyzed in rough order of increasing sophistication.

While a survey of the vast literature relating to the estimation of the parameters in all such constructs might be useful, it would detract from the main purpose of the paper. Hence, only the *a priori* constructs that underlie the econometric testimony are surveyed, thus avoiding such complex areas as model verification, misspecification, and overdetermination.

2. A financial model of the firm

■ Let

$S(t)$ = the value of the firm at the beginning of time t to shareowners of record t.

Then[1]

$$S(t) = \frac{1}{[1+\rho(t)]}[R(t) - O(t) + S(t+1)], \tag{1}$$

where

$\rho(t)$ = the market (or investor) discount rate, i.e., the rate of return investors require for investments of risk comparable to those associated with $S(t+1)$;

$R(t)$ = announced value to be received at the end of t by owners of record t, known with certainty;

$O(t)$ = announced value of all obligations incurred at the end of t, known with certainty; and

$S(t+1)$ = the value of the firm at time $t+1$ to shareholders of record $t+1$, a random variable.

Consider first the use of $\rho(t)$. The "risk-class" approach, which combines time preference and risk aversion into a single measure, is to be compared to the "certainty-equivalent" approach, which separates the two effects.

The "certainty-equivalent" approach hinges upon finding a defensible method of trading off expected performance of a stock against its variation without having to specify the exact personal risk aversion characteristics of each individual stockholder.[2] This approach produces, among other things, a method of evaluating expected performance (mean return) and its variation (variance of that return) that makes minimal use of assumptions about investors' attitudes toward risk. Since, to the authors' knowledge, this approach has been used to date in only two rate cases (and then only tangentially), this paper will start with and primarily emphasize the "risk-class" approach that dominates most current proceedings. Discussion of the alternative will be delayed until after the "risk-class" approach is fully analyzed.

---

[1] Equation (1) is a straightforward application of the familiar capital value equation, in that value is defined to be the discounted stream of cash flows to the owners of the resources less any costs incurred in generating the flows. Thus the equation states that the value of the firm at the beginning of time t is the sum of the net cash flows received during t (arbitrarily assumed to be received at the end of t) plus the value of the firm at the end of t, all quantities discounted to reflect the fact that they lie one period in the future.

[2] Building on one of Tobin's theorems (see [57]), which showed that the optimal mix of risky assets held by an investor (the optimal portfolio problem) could be determined independent of the mix of nonrisky assets (cash) and the aggregate holdings of risky assets (the mean-variance trade-off problem), Sharpe [51], Lintner [30], and Mossin [39] have provided a theory of capital market valuation by combining the necessary conditions for optimal portfolio composition with capital market equilibrium.

In order to assure that $\rho(t)$ will be a constant for all t, as will be done in this paper, it is sufficient to assume that the degree of business, financial, and portfolio risk is unchanged over time.[3] This translates into three requirements: first, that the output market structure does not change in basic demand characteristics; second, that the debt-equity ratio will not change drastically in the future; and third, that investors feel that new investments undertaken by the firm are subject to the same risk that is embedded in past investments.

The assumption of a constant debt-equity ratio permits a simplification of the analysis from a model that explicitly considers debt and equity to one that explicitly considers equity only. If the historical debt-equity ratio is maintained in the financial mix of all future investments, then all present and future interest payments to present and future bond holders can be netted out of the earnings stream, permitting concentration solely upon the return to equity capital.

Consider now $O(t)$, which could have three forms: (1) resources distributed to others besides the owners during t in the form of wages, salaries, and other normal business expenses; (2) the share of current dividends paid to new owners who purchased stock during t; and (3) the present value of future obligations to share the proceeds with such new stockholders. The first form will be submerged beneath the surface of the model by assuming that $R(t)$ is net of any such expenses; the second is assumed to be zero by the convention that any new stock issue is sold "*ex* dividend;" i.e., the first dividend received by new stockholders recruited at the end of t comes out of dividends declared in $t + 1$. Thus $O(t)$ represents solely the present value of the share of the firm's future dividends promised to new stockholders during t; $S(t+1) - O(t)$ is, under this convention, the value of the firm at time $t+1$ to stockholders of record at time t.

---

[3] It is useful to distinguish between three factors that separately influence the choice of $\rho(t)$ first, the inherent riskiness of the market in which the output of the firm is sold, i.e., the "business risk" of the enterprise, second, the "financial risk" that arises out of the capital structure (in particular, the ratio of debt to equity) of the firm, and third, "portfolio risk," which recognizes that a single stock is just one of a number of stock held by investors; since stock prices are interdependent random variables, portfolio risk reflects general economic conditions. Business risk needs no elaboration. Modigliani and Miller (M-M) in [36] specified the conditions under which the value of the firm was independent of its capital structure Since their article appeared, there has been an almost continuous flow of articles specifying more and more exactly where their conclusion fails to hold. Chief among the cases are: (1) when the firm can borrow at an interest rate more advantageous than the individual investor's (Farrar and Selwyn [15]), (2) the presence of a corporate income tax which permits interest to be deducted as an expense (Hamada [22], Litzenberger and Budd [31], Brennan [6], Farrar and Selwyn [15], Myers [40], Robichek and Myers [44], and M-M [35]); (3) acknowledgement of the fact that since the firm has limited liability, individuals would prefer that firms borrow rather than borrow themselves (Sher [52]); and (4) when bankruptcy probabilities of firms are very much affected by the choice of capital structure (Stiglitz [56], Baxter [4], Lerner and Carlton [25], Brewer and Michaelson [7], and Gordon [20])

$R(t)$, the resources distributed to the owners during period t, can be received in two forms: as pure dividends, and/or associated with the issuance of preemptive rights to stockholders. Preemptive rights can either be exercised at the subscription price, in which case they increase the share of $S(t+1)$ owned by stockholders of record at t, leaving $R(t)$ unchanged, or they can be sold to other investors for what they will bring, in which case they enter into $R(t)$ additively with dividends declared. Any earnings retained contribute to increasing $S(t+1)$ but do not enter into $R(t)$.

As several authors have pointed out,[4] under certain assumptions $S(t)$ is invariant with respect to the price of a new stock issue. To see this, suppose that $\dot{N}$ is the number of shares associated with a new issue whose subscription price is P_I; then $\dot{N}P_I = I$ is the total dollars to be raised. Let N be the number of shares outstanding before the issue, and let $S(t+1)$ be the agreed upon value of the firm to new and old shareholders alike, after the investment.[5] Then $S(t+1)/(N+\dot{N})$ will be the price of a share and $S(t+1)/(N+\dot{N}) - P_I$ the market value of a right. Then, ignoring any time lag,

$$S(t) = \dot{N}\left[\frac{S(t+1)}{N+\dot{N}} - P_I\right] + \frac{N}{N+\dot{N}}[S(t+1)] = S(t+1) - I. \quad (2)$$

It is obviously not true that the share price of the firm is invariant with P_I; the importance of this point will be discussed later.

Finally, consider $S(t+1)$, the random variable representing the future value of the firm at time $t+1$ to stockholders of record at $t+1$. Since $S(t+1)$ is a random variable, it is important to specify whose expectations are reflected in it. It certainly should reflect the expectations of those calculating $\rho(t)$; but should it represent the views of present stockholders only, with no regard for the views of new owners? Strictly speaking, a case can be made for excluding new stockholders; but to avoid troublesome complications, the assumption of uniform expectations among all investors, new and old, will be adopted, thus eliminating the problem.

This assumption of uniform expectations should not be treated lightly, since it is a necessary condition for many of the important results in corporate financial theory developed during the last decade. After an early dispute between Miller and Lintner, a classic paper by Lintner, followed by a proof by Robichek and Myers, and a reexamination by Stiglitz and Elton and Gruber,[6] the issues are fairly well established.

Equation (1) will now be applied to the case where the firm is undertaking a continuous expansion program in a taxless world that generates new earnings with a lag from investments financed by both retained earnings and new equity financing. While the possibility exists that a differential between the capital gains and income tax rates might cause a bias in the selection of the sources of capital treated in what follows, such an effect is ignored because of the uncertainty as to what the effective marginal tax rate is for the average investor. (See Brennan[7] for a further discussion.) The Modigliani and Miller (M-M) continuous-financing model[8] is shown to be a special case of this, an important result, since this model underlies most of the models used in regulation.

---

[4] See Beranek [5] and Evans [14].

[5] We have shown this to be a necessary condition. See [12].

[6] See [34], [26], [27], [45], [56], and [13], respectively.

3. The case of continuous financing

■ Consider the following model of continuous financing with stated investment policy. Let:

(1) $X(t)$ = expected earnings from existing assets of record t.

(2) $I_r(t)$ = retained earnings during t with growth rate g_2 such that $I_r(t+k) = (1+g_2)^k I_r(t) \leq X(t+k)$;

$I_s(t)$ = new equity raised by stock sale during t with growth rate g_1 such that $I_s(t+k) = (1+g_1)^k I_s(t)$.

(3) $I(t)$ = announced investment planned for t; $I(t) = I_r(t) + I_s(t)$.

(4) $\beta I_s(t)$ = issue (or flotation) costs for equity sale (in fixed proportion β to dollars raised).

(5) ρ = the single-period rate of return investors require, assumed constant.

(6) r = the single-period expected equity rate of return earned on all investments, assumed constant and known.

(7) m = the number of periods between the time dollars are raised for investment and the time investment begins to generate earnings—the earnings lag.

The announcement of the investment program which will commence at $t + 1$ is made at t; $S(t)$ is the value of the firm at t after the announcement. Period t's dividend is paid at the end of t to shareowners of record t; new shareowners become shareholders of record $t + 1$. Retained earnings $I_r(t)$ and new equity investment $I_s(t)$ are invested at $t + 1$; the investment made in $t + 1$ starts earning in period $t + m + 1$.

Consider first $O(t)$. Initially, the assumption is made that new equity issues are sold at the market price, rather than through the mechanism of a rights issue. As has been demonstrated in (2), setting of P_I equal to the market price in no way changes $S(t)$; $S(t)$ is constant, no matter what the value of P_I. If the new equity issue is sold at the market price, then present investors need promise new investors only their opportunity cost ρ on their investment, assuming perfect markets. Since transactions costs are present, $(1 + \beta)I_s(t)$ dollars of new equity must be sold in order to obtain $I_s(t)$ dollars for investment. Thus the present value of the share promised the stockholders is the discounted sum of their opportunity cost:

$$O(t) = \sum_{j=0}^{\infty} \frac{\rho[I_s(t) + \beta I_s(t)]}{(1+\rho)^{j+1}} = (1+\beta)I_s(t). \qquad (3)$$

---

[7] [6]

[8] Of [37].

Since there are no rights, $R(t)$ is just $X(t)$ less the retained earnings $I_r(t)$.

Finally, since the expected rate of return on new assets is known, $S(t + 1)$ can be computed by standard present value formulas. With these assumptions, equation (2) becomes

$$S(t) = \frac{X(t)}{\rho} + \frac{I_s(t)}{\rho(\rho - g_1)}\left[\frac{r}{(1 + \rho)^m} - \rho(1 + \beta)\right] + \frac{I_r(t)}{\rho(\rho - g_2)}\left[\frac{r}{(1 + \rho)^m} - \rho\right]. \tag{4}$$

Equation (4) holds only at t because of a "start-up" problem caused by the earnings lag. More complicated expressions can be derived for $S(t)$ during the interval between t and the time earnings begin to be generated, and from that time on.[9] For our purposes, equation (4) will be sufficient.

Several observations about (4) are in order. First, note that a sufficient condition for the investment program to increase the value of the firm is

$$\frac{r}{(1 + \rho)^m} > \rho(1 + \beta). \tag{5}$$

Note that if the earnings lag postpones the beginning of the earnings stream far enough, and if flotation costs (β) are high enough, a proposed investment program could decrease the value of the firm to existing owners, even if r, the rate of return earned on assets, exceeded ρ, the investor capitalization rate.

Next, in order to see the relationship between equation (4) and the more familiar models of corporation finance, suppose that there was no timing lag or transaction costs, i.e., $m = \beta = 0$. Further, suppose that instead of I_s and I_r growing at rates g_1 and g_2, they were assumed proportional to $X(t)$, i.e., $I_r(t) = b_r X(t)$ and $I_s(t) = b_s X(t)$. Total investment I is equal to $(b_s + b_r)X(t)$, or just $bX(t)$. Then equation (4) reduces to

$$S(t) = \frac{X(t)(1 - b)}{\rho - rb}, \tag{6}$$

---

[9] The valuation equation for general k has two forms assuming $g_1 = g_2 = g$:

(i) For $0 \leq k < m$

$$S(t + k) = \frac{X(t)}{\rho} + \frac{r[I_r(t) + I_s(t)]}{\rho(\rho - g)}\left[\frac{(1 + \rho)^k - (1 + g)^k}{(1 + \rho)^k}\right] + \frac{I_r(t + k)}{\rho(\rho - g)}\left[\frac{r}{(1 + \rho)^m} - \rho\right] + \frac{I_s(t + k)}{\rho(\rho - g)}\left[\frac{r}{(1 + \rho)^m} - \rho(1 + \beta)\right];$$

(ii) For $k \geq m$

$$V(t + k) = \frac{X(t + k)}{\rho} + \frac{r[I_r(t + k - m) + I_s(t + k - m)]}{\rho(\rho - g)}\left[\frac{(1 + \rho)^m - (1 + g)^m}{(1 + \rho)^m}\right] + \frac{I_r(t + k)}{\rho(\rho - g)}\left[\frac{r}{(1 + \rho)^m} - \rho\right] + \frac{I_s(t + k)}{\rho(\rho - g)}\left[\frac{r}{(1 + \rho)^m} - \rho(1 + \beta)\right].$$

These equations are the solutions of the general model, while equation (4) is the special case of (i) and (ii) when $k = 0$.

which is the familiar equity model of M–M, and, in modified form, of Gordon.[10] Since b_s and b_r do not appear in (6), M–M conclude that value is independent of the level of retained earnings and hence independent of dividend policy.[11] The crucial assumption necessary for this result is that transactions costs be zero, as can be seen by observing that (4) also possesses the property that $S(t)$ is independent of the mix of I_s and I_r if $\beta = 0$. Further, Lintner[12] has argued that the assumption of uniform expectations on the part of all stockholders is necessary for the theorem to hold.

Finally, suppose that the new equity issue did provide for preemptory rights to buy the issue at a price P_I below the market price. What impact will this have? As we have seen, $S(t)$ is unaffected by P_I, the subscription price. To see how this would be true in the case of equation (4), it can be shown[13] that if P_I is expressed as $(1 - \delta)$ times the market price, then (4) can be written

$$S(t) = \frac{X(t)}{\rho} + \bar{v}(t) + \frac{I_s(t)}{\rho(\rho - g_2)}\left[\frac{r}{(1+\rho)^m} - \frac{(1+\beta)\rho}{1-\delta}\right] + \frac{I_r(t)}{\rho(\rho - g_1)}\left[\frac{r}{(1+\rho)^m} - \rho\right], \tag{7}$$

where $\bar{v}(t)$, the present value of all future proceeds from the sale of rights, is defined as

$$\bar{v}(t) = \frac{\delta(1+\beta)I_s(t)}{(1-\delta)(\rho - g_2)}, \tag{8}$$

which reduces to (4) when terms containing δ are combined. It should be emphasized that the market price of the firms' shares is crucially affected by the choice of P_I, since the choice of P_I dictates the number of new shares $\dot{N}$ issued during each period by the relation

---

[10] In [18].

[11] See Hirshleifer [23], M-M [37], Lintner [27], Gordon [19], and Hamada [22].

[12] In [27].

[13] Let the privileged subscription be stated in terms of the right for each existing share to purchase z new shares at $t + 1$ at a price equal to $(1 - \delta)P(t + 1)$. Then

$$v(t) = N(t)z\delta P(t + 1)$$

and in addition

$$N(t)z(1 - \delta)P(t + 1) = (1 + \beta)I_s(t),$$

since $(1 + \beta)I_s(t)$ is the specified required outside equity to be raised. Thus the value of the rights can be expressed in terms of the flotation and discount factors

$$v(t) = \frac{\delta(1 + \beta)I_s(t)}{1 - \delta}$$

For this case the value of future obligations is

$$O(t) = (1 + \beta)I_s(t) + v(t) = \frac{(1 + \beta)I_s(t)}{1 - \delta}.$$

In addition the value of the return, $R(t)$, is now

$$R(t) = X(t) - I_r(t) + v(t)$$

With these changes the valuation equation yields equation (7).

$$\dot{N}(t+k) = \frac{I_s(t+k)(1+\beta)}{P_I};$$

and hence the market price by the relation

$$P(t+k) = \frac{S(t+k)}{N + \dot{N}(t+k)}, \tag{9}$$

where $\dot{N}(t+k)$ is the number of new shares issued during k, and N the number outstanding at the beginning of period $t+k$. This fact requires special care in the use of price data in statistical estimation of the cost of capital, since with rights, time series of prices do not capture all of the returns to stockholders. This point will be discussed in more detail in Section 5.

4. The application to public utilities

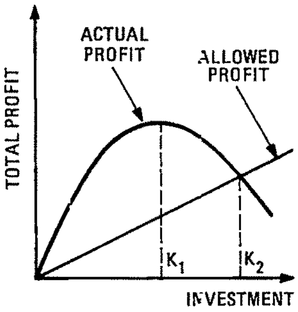
■ All the models presented above were constructed as abstractions of a world of competition, where profit maximization is the assumed goal of the firm; the translation of these models of the finances of competitive firms into formulations useful for rate determination involves some care. First, does the fact that the best the regulated firm can do for its stockholders is to maximize profits subject to the rate of return constraint produce any effects, such as those pointed out by Averch and Johnson (A-J),[14] that prevent the direct application of these models?

This question seems not to have been explicitly raised in the literature; in all instances, the abstractions have been applied by simply assuming that r, the rate of return earned on investment, is the rate of return allowed the utility by the regulatory commissions; thus, the analysis of the regulated firm's financial decisions and its production and input-mix decisions have been carried on in separate watertight compartments.

The main point of the A-J literature is that because of the presence of the rate base, firms tend to adopt projects with higher capital-labor ratios than those for profit maximizing firms, since capital expenses enter the rate base while labor expenses do not. However, this aspect of the problem has not been raised in the financial literature, since factor-mix problems are not expressly considered; all the analysis is carried on in a single-factor world, or a world of fixed-factor proportions.[15] The main impact of the A-J literature on the financial models is that it raises the possibility of overinvestment; i.e., earnings might be retained and/or new equity issues sold for projects that would not be accepted by a profit maximizing firm but would be acceptable to a regulated utility due to the associated increase in the rate base. Figure 1 shows that investment projects totaling k_1 dollars are adopted by profit maximizing firms, while regulated utilities would carry investment to k_2.[16] Besides this possibility, there seem to be no other problems raised by the presence of the rate of return constraint.

FIGURE 1

INVESTMENT LEVELS OF PROFIT MAXIMIZING FIRMS AND REGULATED UTILITIES



---

[14] In [1].

[15] A notable exception is Vickers [61].

[16] Note that this argument is carried out in a single-factor, or a fixed-factor world. See Baumol and Klevorick [2] for a discussion of the necessary conditions for regulated output to exceed unregulated output.

Turning now to more direct and obvious considerations, there are some factors present in the case of public utilities that make the application of financial models of the type described in some ways *more* appropriate than in the case of firms in competitive markets. First, there is the question of the risk of the enterprise. Since public utilities are monopolies, it would be expected that new investments would have the same business risk as those of existing assets. Recent efforts of regulatory commissions to encourage higher debt-equity ratios[17] might require that debt be introduced explicitly into the analysis, if investors feel that regulation does not ensure continued solvency (as many might, following the Penn Central experience). Except for this recent development, the near constant debt-equity ratios shown over time by most utilities make the simplification acceptable.

Another feature of regulated public utilities that permits the separation of debt and equity is that commissions set a *post tax* rate of return; i.e., once the rate is set, taxes are then added in as an expense in establishing revenue requirements. Thus the standard argument[18] that debt-equity ratio considerations cannot be ignored in the presence of corporate income taxes loses its impact. It is interesting to note that this issue was the central point of an exchange between M-M and Gordon[19] over the appropriateness of M-M's test of their debt-equity theorem using data from the electric power industry. In that article, M-M claim that their results tend to support their debt-equity theorem; however, as Gordon points out, "only if their (M-M's) sample had consisted of nonregulated industrial companies would their results have been evidence in support of their theorem."[20] A recent article by Elton and Gruber[21] has reopened this point by claiming that Gordon has incorrectly specified the argument.

Another factor that makes the models more acceptable in public utility applications than in more competitive situations is that since r, the allowed rate of return set by the regulatory commissions, is public knowledge, investor expectations about the future earnings of the utility should show a smaller variance than expectations about nonregulated firms, thus making the assumption of uniform expectations in the financial models more acceptable. The rather heavy use of the equity market by most utilities and the pressure of a substantial lag between the time investment dollars are obtained and the time such investments begin to generate earnings require that the financial models contain these characteristics; simple formulations where no new equity financing is allowed,[22] or where no earnings lags are introduced, are not too useful.

Finally, a common occurring theme in regulatory models is that the rate of return should be specified equal to the cost of capital. The framework is that of the single-period rate of return and the single-period investor discount rate. In the M-M world as depicted by equation (6), this would leave the value of the firm unchanged. A frequent argument is that this should be the case since $r - \rho$, in a sense, represents economic unit profit in this frictionless, instantaneous world. To see how this can be misleading in the case where $m \neq 0, \beta \neq 0$, let the ratio of new equity financing to total investment be equal to α, then:

---

[17] See Hearing on the A T &T Interim Revision and Order [60].

[18] See M-M [35].

[19] See [38], [21], and [8].

[20] [21], p 1271.

[21] [13].

[22] Such as those analyzed by Gordon [20] and Lintner [28].

$$\begin{aligned} I_r(t) &= (1 - \alpha)I(t) \\ I_s(t) &= \alpha I(t), \end{aligned} \tag{10}$$

with $\alpha \leq 1$.

With this specification, the r such that $S(t)$ remains unchanged is

$$r = \rho(1 + \rho)^m[1 + \alpha\beta], \tag{11}$$

a relationship reflecting explicitly the parameters m, α, β, and ρ. What this condition specifies is the equating of "actual" internal rate of return to "actual" cost of capital to the utility. This requires explicit recognition of the institutionalized characteristics of the corporation (r as a function of ρ, m, α, and β) where

$\dfrac{r}{(1 + \rho)^m}$ = "actual" internal rate of return

and

$\rho(1 + \alpha\beta)$ = "actual" cost of capital to the utility for the specified investment program.

This appears to be closer to the intent of rate of return specification by regulation than that of simply requiring the single-period rate of return to equal the capitalization rate.

5. The use of valuation models in regulatory hearings

■ While this paper is not intended to cover all facets of rate of return regulation, it does seem appropriate to preface this section on the use of financial models in rate hearings with a few words on the general setting. More complete discussions are available in recent texts on regulation such as that of Phillips or, especially, Kahn, upon which much of this discussion is based.[23]

Prior to the *Hope* case in 1944,[24] the regulatory commissions by and large followed in their deliberations the dictates of the *Smith* v. *Ames* decision[25] that the owners of utility stock were entitled to a "reasonable compensation"; such a compensation was fixed by applying a fairly standard (and noncontroversial) rate of return to a rate base that was to reflect "fair value," a highly controversial concept. While *Smith* v. *Ames* mentioned four possible criteria by which a "fair value" base might be determined, one, the present cost of construction, dominated. Consequently, commission hearings were filled with testimony from engineers, accountants, economists, and the like, all directed at deriving what was the "correct" current reproduction cost of a utility's capital stock. There then followed a period of private deliberation by the commission (unkindly, but not without foundation, called a "trance" by many authors) from which emerged the correct fair value of the rate base. The *Hope* decision reversed *Smith* v. *Ames* by arguing that one method of determining fair value mentioned in *Smith* v. *Ames,* the market value of stocks and bonds, resulted in a hopelessly circular argument, in that the chief determinant of market value is anticipated earnings, whose levels are governed by what fair value the commissions finally decide upon. To have fair value depend upon fair value was of little help, they concluded.

---

[23] [43] and [24], respectively.

[24] Federal Power Commission v. Hope Natural Gas Company, 320 U.S. 591 (1944).

[25] Smith v. Ames, 169 U.S. 466 (1898).

While the *Hope* decision has been characterized as a withdrawal by the courts from the setting of any formal criteria, it did suggest that the ultimate outcome of the regulatory hearings should not impair the ability of the utility to attract capital. This requirement, plus the 1923 *Bluefield*[26] dictum that owners were entitled to returns equal to those found in firms of comparable risk, forms the basis for most recent deliberations. Since capital attractiveness and comparable risk are concepts dealing with the rate of return to capital, the years since the *Hope* case have seen a gradual decrease in controversy over what the rate base should be; most jurisdictions now require that it be calculated on an original cost basis. Almost all of the controversy has shifted to the choice of a proper rate of return. A whole new industry has sprung up devoted to determining the "fair" rate of return, while the "fair rate base" industry languishes in its old age; technological progress takes its toll. Quite naturally, the recent wave of quantitative reform in the field of corporate finance has had its impact upon the hearings; the various valuation models described below have found their way into the testimony and have been proposed as evidence by regulators and regulatees alike in recent years.

Several comments are in order before proceeding. First, a public utility's cost of capital should be independent of commission decisions since the opportunity cost is not changed, thus removing the circular problem that led to the demise of *Smith* v. *Ames*. The price of a utility's stock will change with commission decisions, by exactly the argument presented in the *Hope* case: Commission decisions affect earnings which in turn affect the stock price, leaving the cost of capital or the investor discount rate unchanged. However, if commissions take actions that affect the risk of the utility (as, for instance, actions that would lead to changing the debt-equity ratio), they lay themselves open to exactly the circularity criticisms that finished off *Smith* v. *Ames*.

But there is more to the argument. The rapidity with which stock prices adjust to changes in expectations about earnings guarantees that individuals holding utility stocks will always earn only the fair rate of return that commissions seek; thus commission decisions affect only the capital gains or losses of those holding the stock of the utility at the time of the decision. If they incorrectly estimate the cost of capital, the market price will signal the error by an appropriate adjustment, thus providing a partial *ex post* test of the correctness of the commission decision. (It is only a partial test, because of the rate of return's interaction with the rate base, and the confounding of the "cost of capital" and growth allowances [see equation (19)].) This is a rather disconcerting fact, since determining what is a "fair" capital gain or loss involves the historical price paid by the investor, and what price appreciation expectations he had, leading us dangerously near the circularity argument again. This latter point deserves repetition: while it is true that the pure cost of capital is not affected by commission decisions (subject to the risk proviso) *commission deliberations on proper growth allowances have exactly the same circularity faults* as in the "fair rate base" case. Growth of earnings, like earnings, is governed by commission decisions; commission decisions cannot be made to depend upon commission decisions.

---

[26] Bluefield Water Works & Improvement Company v. Public Service Commission of West Virginia, 262 U S. 679, 693 (1923)

The "capital attractiveness" criteria warrants specific comment. As equation (5) illustrated, since new investors must always be promised the cost of capital as a necessary condition for purchase because of the near perfect nature of capital markets, any new equity issue can depress the value of the firm to its present owners if the allowed rate is less than the cost of capital, adjusted for transactions costs and lags. Thus utilities can always attract capital by reducing the price of their stock until new investors feel that they will obtain the market rate, but only at the expense of existing owners, thus raising the question of the utilities' willingness to raise new capital.[27] The requirement that utilities provide sufficient new capital to insure that an adequate level of service be provided to the consumer coupled with an allowed rate of return (however figured) below the cost of capital will result in capital losses to present owners if the new capital must be raised in equity markets.

It is obviously impossible to do justice to all aspects of the problem in such a limited space; the points discussed are only the subset of the regulatory issues that play a role of particular importance when valuation models are introduced into legal proceedings.

6. The models themselves

■ The oldest and perhaps best known of the rate of return models is that associated with the "comparable earnings test." The hypothesis underlying the test is that firms can be partitioned into groups whose risks are similar, taken to mean that the risk-adjusted discount rates for all firms in a given risk class are identical. The investor discount rate is obtained by dividing the current earnings per share by the current price, implying that price is equal to capitalized earnings. The valuation model underlying the concept is then that price is equal to the present value of a constant earnings stream discounted at the rate obtained by dividing earnings by price. Thus, if P is price and E earnings per share, then

$$P = \sum_{t=1}^{\infty} \frac{E}{(1 + \rho)^t} = \frac{E}{\rho}; \qquad (13)$$

---

[27] This is not a new idea; both Kahn [24], p. 46, and Scherer [50], pp 524–27, present numerical examples of this phenomenon.

hence, ρ is taken to be measured by the earnings-price ratio. Frequently, the model appears in the form which will now be described.[28]

From a group of companies, deemed comparable in risk characteristics, the approach determines the mean (average) rate of return to equity, r, equal to P/B, and the mean (average) market to book ratio, P/B, over some specified time period, where

E = earnings per share,
P = market value of company (share price)
and
B = book or equity value per share of company.

The key assumption required in this test is that the market to book ratio is a constant proportion of single-period rate of return for all companies in this selected "risk class," that is,

$$(P/B)_i = Cr_i \text{ for all companies } i, \tag{14}$$

where C is a proportionality constant, seen to be a constant function of the investor discount rate in (13).

It is argued that a utility will be attractive to investors who prefer this particular "risk class" of companies only if

$$(P/B)^*_{\text{utility}} = (P/B),$$

where

$(P/B)^*_{\text{utility}}$ = the desired market to book ratio for the utility (without the asterisk it denotes current quantity),

and

(P/B) = the "average" market to book ratio for the comparable group.

Using equation (14) it follows that

$$\begin{aligned} r^*_{\text{utility}} &= \frac{1}{C}(P/B)^* \\ &= \left[\frac{r_{\text{utility}}}{(P/B)_{\text{utility}}}\right](P/B)^* \qquad (15) \\ r^*_{\text{utility}} &= \left[\frac{(P/B)}{(P/B)_{\text{utility}}}\right]r_{\text{utility}}. \end{aligned}$$

Thus equation (15) indicates how the utility's rate of return to equity needs to be adjusted for comparability in terms of measures which are statistically observable. The crucial assumption, as previously indicated, is that market to book is a constant proportion of rate of return.

The first observation is in regards to the variable specification generally used. Price data are obviously not suitable for this analysis if equity financing has involved privileged subscription (rights). As the previous analysis has shown [equation (9)], the price data will be critically affected by the discount for privileged subscription. When rights are issued, the value of the firm to its stockholders is no longer reflected solely in the price of the stock: a portion of the value is transferred to the right. A better variable specification would be the total value approach using

---

[28] This formulation was demonstrated to us by P. B. Robinson [47] of AT&T and will follow that of Robinson's, using the notation of this paper

S = total value of the Company to its stockholders
B = total book value or equity of Company
and
$(S/B)_i = Cr_i$ for all companies i. (16)

This removes any privileged subscription effects and is in the same spirit of this general approach.

Assuming the variables are specified as indicated, equation (4) can be used to analyze this formulation. Suppose $X(t) = rB(t)$; then (4) can be written

$$\frac{S(t)}{B(t)} = r\left\{\frac{1}{\rho} + \frac{I_s(t)}{B(t)\rho(\rho - g)}\left[\frac{1}{(1+\rho)^m} - \frac{\rho(1+\beta)}{r}\right] + \frac{I_r(t)}{B(t)\rho(\rho - g)}\left[\frac{1}{(1+\rho)^m} - \frac{\rho}{r}\right]\right\}. \tag{17}$$

For equation (17) to satisfy equation (16) the quantity in { } must be constant and independent of r. Equation (17) clearly indicates that this is not true, even if the simplifying assumption is made that historical earnings $X(t)$ are given by multiplying book value by r. It is true if either

(1) $I = 0$, i.e., no growth,

or

(2) $m = 0$, $\beta = 0$, and $r = \rho$.

If either of the above holds, then

$$\frac{S(t)}{B(t)} = \frac{1}{\rho}r. \tag{18}$$

Alternative (1) leads to the conclusion that the comparable earnings test is a static or no growth formulation, a fact widely recognized. However, for this static case the formulation does what it intended to do, that is, construct a mechanism for allowing ρ, the capitalization rate (and a function and specifier of risk class) to be used in rate of return calculations. It is of interest to note that it will not be easy to "modify" this construct to handle the growth cases.

Alternative (2) has provided the inspiration for two authors, Gelhaus and Wilson,[29] to attempt a rescue of the comparable earnings approach. If the rate of return allowed is equal to the cost of capital and there are no time lags or flotation costs, then a firm can grow by retaining earnings and/or obtaining new equity financing without changing the ratio of value to book. Thus, even with growth, commissions could obtain an estimate of the cost of capital using earnings-price ratios.[30] Unfortunately, the cure is worse than the disease; in order to set the rate of return to the cost of capital, the authors propose an iterative process which forces the price of the stock down (and the earnings/price ratio up) until equality appears between r and ρ. The problems raised by this tax upon the present owners plus the adverse effect upon incentives to improve performance (in that any achievement by the firm that increases stockholder expectations and hence price is followed immediately by a reduction in the allowed rate)[31] make the plan appear unacceptable, even though it does guarantee the commission a direct estimate of the cost of the capital. Equation (17) clearly indicates that institutionalized characteristics, m, β, and g, as well as $B(t)$ cannot be "solved" out of the "constant" part; hence there will *never* be a proportionality constant, C, unless every firm is identical.

---

[29] [17].

[30] Kahn [24], p 49, in his review of Gelhaus and Wilson [17] seems to think that the argument is valid only if growth is financed out of retained earnings. This is not the case, as equations (4) and (17) demonstrate.

The problem of quantitatively treating growth in rate of return models has been of recent central concern.[32] One approach has been to use discounted cash flow (DCF) to determine ρ and then set the rate of return, r, equal to it. A number of conceptual difficulties with this approach have been well documented and will not be discussed here;[33] this approach will be analyzed in the framework of the model of this paper. The DCF method can be written as

$$\rho = \frac{d}{P} + g, \tag{19}$$

where

d = dividends per share,
P = share price, and
g = "anticipated" growth in dividends per share.

Equation (29) comes directly from the conceptual construct which yielded equation (1).[34] The approach is that d/P is the current (anticipated) dividend yield, both observable and computable. The g term is the "anticipated" growth in dividends which the investor expects this investment to yield (over an infinite time horizon). The g term is *not* directly observable and must be estimated from surrogate data. This expectation parameter is characteristically estimated from historical price data, dividend data, or earnings data.[35] In any one of these cases, any return due to past or anticipated privilege subscription (rights) should be included, if past price data are used. More importantly this formulation implicitly assumes that the growth term, g, is independent of the other variables, in particular, share price, P. If growth of the firm is achieved by issuing new shares, then equation (19) is an incorrect specification of the cost of capital, even if the simplifying assumptions of the M-M world are assumed. To see this, consider equation (6), the M-M simplification of equation (4).

---

[31] See Ross [49].

[32] See Testimony of Myers [59], Testimony of Gordon [58], Colbert [10], and Sparrow [55].

[33] See Testimony of Myers [59] and Roseman [48]

[34] See McDonald [32] and Testimony of Myers [59] in particular for discussion of this formulation.

[35] See Testimony of Myers [59] for a recent example.

After dividing both sides of equation (6) by N, the current number of shares outstanding, and solving for ρ, we obtain

$$\rho = \frac{x(t)(1 - b_r)}{P(t)} + r\left[b_r + b_s\left(1 - \frac{B(t)}{P(t)}\right)\right],$$

where

$$x(t) = \frac{X(t)}{N}$$

and where the substitution $b_r + b_s = b$ has been made. If b_s is zero, then $X(t)(1 - b_r)$ is in fact $d(t)$ and rb_r is in fact the growth in dividends per share. If b_s is not zero, i.e., if there is reason to believe that growth will be financed by new stock issues, the independence of g from $P(t)$ cannot be assumed. The argument then involves determining what is "the maximum acceptable level" of new equity issues that permits ignoring their effect on g and continuing the use of (19). It should be clear that if (19) is unacceptable in the M-M world of (6), it is even more unacceptable if the conditions of (4) are assumed to hold.

The above argument serves as a good introduction to the application of a more sophisticated model that allows for both retained earnings and new equity issues as sources of funds. In 1963, A. R. Colbert[36] suggested a model that would yield that rate of return on equity necessary to allow dividends to grow at a fixed rate of i percent per year, when it was assumed that the utility issued new shares at a rate that produced an r percent increase in total shares per year. Net proceeds to shares also grew at i percent per year holding the payout ratio (dividends to earnings) constant. Sparrow[37] demonstrated that if a valuation equation of the form of equation (19) was added, then the Colbert and M-M models became identical.

It can be shown that the formulation of Gordon,[38] which explicitly allows for financing by new equity and retained earnings, is also equivalent to the M-M formulation.[39] The Gordon "extended" DCF model is

$$P(0) = \frac{(1 - b)rE(0)}{\rho - br - vq}, \qquad (21)$$

---

[36] In [10].

[37] In [55].

[38] [18], p. 120.

[39] The identity with the M-M model can be shown by observing that M-M in [37] use the parameters

k_r = retention rate as a fraction of *earnings*; and

k_e = stock financing rate as a fraction of *earnings*;

while Gordon in [18] uses

b = retention rate as a fraction of *earnings*; and

q = stock financing rate as a fraction of *book value.*

Thus the difference is only in parameter specification, i.e.,

$$b = k_r \text{ and } q = rk_s$$

Otherwise, their continuous growth models with outside equity financing are identical. (Gordon uses price per share whereas M-M use total value; however, the equivalence is direct) To show this equivalence, consider the Gordon model. Share price is defined as where b is the retention rate, $E(O)$ the current book value, r the return on book, (thus $rE(O)$ is earnings), ρ the investor discount rate, q the fraction of book value raised by new stock issues, and v the "fraction of the stock issue . . . that accrues to the (present) stock holders."[40]

As was demonstrated, the principal improvement over the DCF method outlined previously is that the sources of dividend growth are specified exactly, rather than submerged behind an overall growth rate in dividends per share, g. In contrast to the DCF method, the Colbert-Gordon-M-M model assumes specific rates of growth for retained earnings and equity financing. This advance is not to be taken lightly; however, even this model has been shown to lack important characteristics of the real world as spelled out in the discussion preceding equation (4), and hence should be used only with extreme care, if at all, in cases where lags, transactions costs, or any of the other factors assumed away to obtain (4) are known to be significant.

Finally, returning to the point raised in the Gelhaus-Wilson discussion above,[41] this model collapses into the simple comparable earnings model when the rate of return equals the cost of capital. To see this, observe that if r is set equal to ρ in equation (6), the M-M equivalent of (20), the $(1 - b)$ term cancels out of the numerator and denominator, leaving (6) identical in form to equation (13), the comparable earnings equation.

All of the models discussed above share the common defect of ignoring the impact of a change in the debt-equity ratio. As of this writing, the only model known that deals directly with this problem is the normative construct underlying the econometric model of Gordon.[42] While Gordon is imprecise in the specification of the valuation model used, examination of the arguments surrounding

---

$$P_0 = \int_0^\infty (1 - b)rE(t)e^{-\rho t}\,dt$$

and $E(t)$, book value per share, is defined in terms of a differential equation·

$$\frac{dE(t)}{dt} = rbE(t) + vqE(t),$$

where v is defined as the fraction of investment at t accruing to the existing shareholders. Solving the above yields

$$\frac{P_0}{E_0} = \frac{(1 - b)r}{\rho - br - vq}.$$

Because of v it is not directly obvious that this result is equivalent to M-M Recognizing that

$$v = 1 - \frac{E(t)}{P(t)}$$

and after some manipulation, we arrive at

$$\frac{P_0}{E_0} = \frac{r(1 - b) - q}{\rho - rb - q},$$

which is equivalent to the result of [18]

[40] [18], p. 120.

[41] Pp. 557–58 *infra*.

[42] Testimony of Gordon [58].

the econometric model suggests that the model is a modification of equation (21) to include effects on value arising from:

(1) The issuance of new debt,
(2) The presence of old debt in the capital structure, and
(3) The possibility that some of the parameters of the model (ρ and r in particular) are functions of the decision variables.

According to Gordon, interest on new debt, i, is a function of the debt-equity ratio, h, and the investment rate, g. If i is affected by h and g, r, the return to equity, will be affected as well since r is calculated as a residual after deduction of interest. Further, ρ is assumed affected by b, the retention ratio, h, the debt-equity ratio, and q, the stock-financing rate. Thus equation (21), rewritten to reflect these dependences, would become

$$P(0) = \frac{(1 - b)r(h,g)E(0)}{\rho(b,h,q) - br(h,g) - qv}. \tag{21a}$$

The explicit consideration of h, the debt/equity ratio, and its impact on ρ and r is a definite improvement over the preceding models: previous arguments indicated the validity of these assumptions. Further, the assumption that the rate of equity financing, q, might depress the stock price (increasing ρ) can be supported by arguing that in a real world of diverse expectations, additional investor interest may only be generated by lower prices—a consequence of the familiar downward sloping demand curve mentioned in the discussion of $S(t + 1)$ in equation (1). Note that equation (4) clearly demonstrates that with *flotation costs* (not assumed by Gordon) equity financing increases the firm's cost of capital. However, if privileged subscription is used and share price movement is made a surrogate measure to reflect changes in capital costs, equity issues could take on the appearance of increasing costs when, in fact, the introduction of "rights" could have a tendency to decrease actual costs of equity subscriptions (a low issue price might make the issue attractive) even though price could change dramatically.

The assumed relation between ρ and b, the retention rate, is on much less firm ground. The claim that risk changes when retained earnings are substituted for new equity financing (thus reducing dividends) while leaving the investment policy unchanged is as yet without foundation. Perhaps it is an artifact left over from the earlier Gordon—M-M dispute over the impact of dividend policy upon valuation, during which the case of a change in the dividend policy with no change in investment policy was constantly confused with changes in dividend *and* investment policy.

Equation (4), the continuous growth model with assumptions similar to those of Gordon,[43] demonstrates that earnings lag simply postpones into the future earnings from investment. This is equivalent (with a constant payout ratio) to postponing dividends into the future. Value and price, of course, decrease as m increases, but this should not be taken as an indication of increased riskiness. What is required is a correct evaluation of "actual" internal rate of return versus cost of capital, not the single period rate of return.

---

[43] In [18].

It should be noted that the problems discussed in regard to the DCF method are equally applicable here unless the instantaneous, frictionless world of M-M[44] is assumed.

7. The certainty equivalent approach of Sharpe-Lintner-Mossin

■ While the approaches of Sharpe, Lintner, and Mossin[45] cannot be strictly regarded as perfect substitutes for one another, their differences are of a technical nature, not of central concern to the development here. The formulation taken will be cast in terms of the variables used in the preceding section so direct comparisons can be made. The following standard assumptions are made in the three capital asset pricing models:

1. All markets are competitive in that prices adjust instantaneously to clear all markets.
2. There exists a risk-free asset yielding a certain interest rate r^* per period.
3. Investors have uniform expectations of the dividends and price appreciation for each stock.
4. Investors are risk averters but not to the extent of using "safety first" criteria. Portfolios are optimized on the basis of mean-variance characteristics.

With these assumptions it can be shown[46] that $S(t)$, the value of the firm at the beginning of period t, can be expressed as

$$S(t) = \frac{1}{1 + r^*}\left[R(t) - O(t) + \bar{S}(t + 1) - W\right]$$

or (22)

$$S(t) = \frac{1}{1 + r^*}[\bar{H} - W],$$

where

1. $R(t)$ and $O(t)$ are as defined in equation (1), and $\bar{S}(t + 1)$ is the *expected* value of $S(t + 1)$, uniformly held by all investors,
2. r^* = the risk-free interest rate,
3. $\bar{H} = R(t) - O(t) + \bar{S}(t + 1)$; and
4. $W = C^* \sum_j \operatorname{cov}(H, H_j)$, where C^* is the "price of a dollar of risk," a constant for all firms, and $\sum_j \operatorname{cov}(H, H_j)$ is the sum of the covariances of H with all other firms' cash flows. Thus W is the dollar risk adjustment for the firm under consideration.

While the modifications introduced into equation (1) to yield equations (4) and (7) could be introduced directly into equation (22), it is not necessary to go to the more complicated version to compare the results. A comparison of equations (1) and (22) shows that, in general, one cannot express $S(t)$ as $\bar{H}$ discounted by a risk-adjusted interest rate that is independent of $\bar{H}$, i.e., the expectation on H. This can be seen by observing that equation (22) can be written as

---

[44] See [37]

[45] In [51], [30], and [39], respectively

[46] See Lintner [30]

$$S(t) = \frac{\bar{H}}{1+\gamma},$$

where

$$\gamma = \frac{\bar{H}r^* + W}{\bar{H} - W};$$

hence γ, the risk-adjusted rate, is dependent on $\bar{H}$. If, however, there exists a relationship between $\bar{H}$ and W, or more fundamentally between $\bar{H}$ and $\text{cov}(H,H_j)$ such that $\text{cov}(H,H_j) = \alpha\bar{H}$, then

$$\gamma = \frac{r^* + \alpha}{1 - \alpha},$$

which is independent of $\bar{H}$, the expected return of the firm under consideration.

Thus if there exists a class of firms whose $\bar{H}$ and $\text{cov}(H,H_j)$ are similar, then the Sharpe-Lintner-Mossin model can be couched in terms of the more traditional "risk-class" approach, a fact pointed out by Hamada and others in a different context.[47] It is of interest to note that this idea has recently been used somewhat tangentially in two rate cases, involving the testimony of Friend and Myers,[48] in a manner analogous to this specialization. In each case the approach has been to use the risk measure (the terms containing covariances) to demonstrate comparability of a collection of firms in terms of risk. The testimony of Friend, in particular, demonstrates this approach to find the risk class to test comparability in applying the comparable earnings test. Equation (22) can be put into the more "familiar" form used by Myers and Friend in their testimony by expressing C^*, the price of a dollar of risk, as[49]

$$C^* = \left(\frac{R_m - r^*}{\sigma_m{}^2}\right) \Big/ \sum_j S_j(t), \tag{23}$$

where

R_m = the expected rate of return on the "optimal" portfolio (traditionally the entire market),
σ_m^2 = the variance of that "optimal" portfolio, and
$\sum_j S_j(t)$ = the value of all stocks traded in the market.

The choice of approach to rise in rate cases (the risk class or certainty equivalent) need not be decided on the basis of which has the most relevant assumptions. The two approaches can be considered equivalent, as shown by Hirshleifer, Hamada, and others, if one is generous in his interpretation of the term, "risk class," to mean a class of firms whose earnings random variables are related by a constant such as α in the equation for γ above.[50] In particular, they have been shown to be equivalent to the M-M model discussed earlier.[51] The above approach was used to demonstrate the assumptions implicit in any use to demonstrate "comparability."

---

[47] In [22].

[48] See [41] and [59].

[49] See [42], p. 70.

[50] See [23] and [22], respectively.

[51] See [37].

Thus while this approach has an elegance superior to the more simple-minded "risk-class" approach and may in fact be more amenable to empirical application, it should not be considered an independent alternative in rate of return applications.

8. Conclusions

■ In conclusion it has been shown that all models considered have assumed either explicitly or implicitly the financial market conditions used in the development of the M-M model. Accordingly, the consequences, criticisms, and attributes of the M-M financial world seem to represent the state of current rate of return modeling.

The central theme in all of the rate of return models has been toward establishing that relationship which allows the "regulated" rate of return (single-period) to be specified as equal to the utility's cost of capital. This is a task of immense difficulty for more reasons than apparent in the above models and analyses. Some work has begun in these areas and is indicative of where future research in this area is moving.

First, cost of capital is an investment criterion applied at the margin (a marginal "cost"), whereas "fair" or allowed rate of return is an "equity" criterion applied uniformly (an average "cost"). The importance of understanding this difference cannot be too strongly emphasized.[52] It is here where most confusion and misapplication resides. One central issue of future research must be to establish quantitative relationships between the two, in the "average" equal to "marginal" sense, for this relationship is one of the infrastructural elements in rate of return modeling.

In addition, cost of capital under conditions of growth requires a dynamic analysis in a context of optimization. The present models are static and contain no explicit optimization.[53] The future perspective should be one that allows the time development of the utility's capital outlays to be analyzed directly in a context of optimization as an interdependent chain of decisions to satisfy an evolution of demand. It is here that the equating of rate of return with cost of capital to insure desired growth will be most meaningful.

Finally, the framework of rate of return analysis needs to be broadened so that the surrogates for the competitive market mechanisms as well as the interaction of regulatory requirements can be directly accounted for. Far too often in the area of regulatory analysis, the general approach has been, to use a medical analogy, like that of a physiologist: explaining the meaning of the separate parts and alluding to the whole. Unfortunately, the outcome of this approach has quite often been analogous to that of a pathologist: doing post mortems in search of diseased tissue. Perhaps a better approach would be to adopt the posture of a general practitioner (which is also regaining favor in the medical profession), where the individual (utility) is viewed in his (its) total environment, and questions such as how much oxygen (revenue) is needed to sustain life in the left toe (one service offering) is not seriously considered as the primary aim of public health (regulation).

---

[52] See Solomon [53] for a recent discussion

[53] See Davis [11] for a recent example of a dynamic analysis with optimization

References

1. Averch, H. and Johnson, L. L "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol 52, No 5 (December 1962), pp. 1062–69.

2. Baumol, W. J. and Klevorick, A. K. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–90.

3. Baumol, W. J. and Malkiel, B. "The Firm's Optimal Debt-Equity Combination and the Cost of Capital." *Quarterly Journal of Economics*, Vol. 81, No. 325 (November 1967), pp. 547–78.

4 Baxter, N. "Leverage, Risk of Ruin, and the Cost of Capital." *Journal of Finance*, Vol. 22, No. 4 (September 1967), pp. 385–404.

5. Beranek, W. *Analysis for Financial Decisions.* Homewood, Ill.. Richard D. Irwin, Inc., 1963, pp. 209–16.

6. Brennan, M. J. "Taxes, Market Valuation and Corporate Financial Policy." *National Tax Journal*, Vol. 23, No. 4 (December 1970), pp. 417–27

7 Brewer, D. E. and Michaelson, J. B. "The Cost of Capital, Corporation Finance, and the Theory of Investment: Comment." *The American Economic Review*, Vol. 55, No. 3 (June 1965), pp. 516–23.

8. Brigham, E. and Gordon, M. J. "Leverage, Dividend Policy, and the Cost of Capital." *Journal of Finance*, Vol. 23, No. 1 (March 1968), pp 85–104

9. Chen, H. "Valuation under Uncertainty " *Journal of Financial and Quantitative Analysis*, Vol. 2, No. 3 (September 1967), pp. 313–25.

10. Colbert, A R. "Common Stock Earning Allowances for Utilities as Adjusted for Growth." *Land Economics*, Vol. 39, No. 1 (February 1963), pp. 45–52.

11 Davis, B. E. "Investment and Rate of Return for the Regulated Firm." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 245–70.

12. ——— and Sparrow, F T. "Equity Issues, Imperfect Markets, and Diverse Expectations." Mimeographed 1971.

13. Elton, E. and Gruber, M. "Valuation and the Cost of Capital for Regulated Industries." *Journal of Finance*, Vol. 26, No. 3 (June 1971), pp. 661–70.

14. Evans, G. H , Jr. "The Theoretical Value of a Right." *Journal of Finance*, Vol 10, No 1 (March 1955), pp. 55–61.

15. Farrar, D. E. and Selwyn, L. L. "Taxes, Corporate Financial Policy, and Return to Investors " *National Tax Journal*, Vol. 20, No. 4 (December 1967), pp. 444–54.

16. Fisher, I. *The Theory of Interest.* New York: Macmillan Co., 1930.

17 Gelhaus, R. J. and Wilson, G. D. "An Earnings Price Approach to Fair Rate of Return in Regulated Industries." *Stanford Law Review*, Vol. 20, No. 2 (January (1968), pp 287–317.

18. Gordon, M. J. *The Investment, Financing, and Valuation of the Corporation.* Homewood, Ill.. Richard D. Irwin, Inc., 1962.

19. ——— "Optimal Investment and Financing Policy." *Journal of Finance*, Vol. 18, No 2 (May 1963), pp. 264–72.

20 ———. "The Savings Investment and Valuation of a Corporation." *The Review of Economics and Statistics*, Vol. 44, No. 1 (February 1962), pp. 37–51.

21. ———. "Some Estimates of the Cost of Capital to the Electric Utility Industry, 1954–57: Comment." *The American Economic Review*, Vol. 57, No. 5 (December 1967), pp. 1267–78

22. Hamada, R. S. "Portfolio Analysis, Market Equilibrium, and Corporation Finance." *Journal of Finance*, Vol. 24, No. 1 (March 1969), pp. 13–32.

23. Hirshleifer, J. *Investment, Interest, and Capital.* Englewood Cliffs, N.J.. Prentice-Hall, Inc., 1970.

24. Kahn, A. E. *The Economics of Regulation*, Vol. 1. New York: John Wiley & Sons, Inc., 1970.

25. Lerner, E. and Carlton, W. "The Integration of Capital Budgeting and Stock Valuation." *The American Economic Review*, Vol. 54, No. 4 (September 1964), pp. 683–702.

26. Lintner, J. "The Cost of Capital and Optimal Financing of Corporate Growth." *Journal of Finance*, Vol. 18, No. 2 (May 1963), pp. 292–310.

27 ———. "Dividends, Earnings, Leverage, Stock Prices, and the Supply of Capital to Corporation." *The Review of Economics and Statistics*, Vol. 44, No 3 (August 1962), pp. 243–69.

28. ———. "Optimal Dividends and Growth under Uncertainty " *Quarterly Journal of Economics*, Vol. 78, No. 314 (February 1964), pp. 49–95.
29. ——— "Security Prices, Risk, and Maximal Gains from Diversification." *Journal of Finance*, Vol. 20, No. 5 (December 1965), pp. 587–613.
30. ——— "The Valuation of Risk Assets and Selection of Risky Investments in Stock Portfolios and Capital Budgets." *The Review of Economics and Statistics*, Vol. 25, No 1 (February 1958), pp. 65–86.
31. Litzenberger, R. H. and Budd, A. P. "Corporate Investment Criteria and the Valuation of Risk Assets." *Journal of Financial and Quantitative Analysis*, Vol. 5, Nos. 4 & 5 (December 1970), pp. 395–417.
32. McDonald, J. G. "Required Return on Public Equities: A National and Regional Analysis, 1958–1969." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 2 (Autumn 1971), pp. 503–14
33 Markowitz, H. "Portfolio Selection." *Journal of Finance*, Vol. 7, No 1 (March 1952), pp. 77–91.
34. Miller, M. "Management of Corporate Capital. Discussion." *Journal of Finance*, Vol. 18, No. 2 (May 1963), pp. 311–18.
35 Modigliani, F. and Miller, M. H. "Corporate Income Taxes and the Cost of Capital: A Correction." *The American Economic Review*, Vol. 53, No. 3 (June 1966), pp. 333–91.
36. ———. "The Cost of Capital, Corporation Finance, and the Theory of Investment." *The American Economic Review*, Vol. 48, No 3 (June 1958), pp. 261–97.
37. ———. "Dividend Policy, Growth, and the Valuation of Shares." *Journal of Business*, Vol. 34, No. 4 (October 1961), pp. 411–33.
38. ——— "Some Estimates of the Cost of Capital to the Electric Utility Industry, 1954–57." *The American Economic Review*, Vol 56, No. 3 (June 1966), pp. 333–91.
39. Mossin, J. "Equilibrium in a Capital Asset Market." *Econometrica*, Vol. 34, No. 4 (October 1966), pp 768–83.
40. Myers, S C. "Taxes, Corporate Financial Policy, and Return to Investors: Comment." *National Tax Journal*, Vol. 20, No 4 (December 1967), pp 455–62.
41. New York Public Service Commission. *New York Telephone Co* Prepared Testimony, I. Friend. N.Y.S C Docket 26080, 1971.
42. Peterson, D. E. and Haydon, R. *A Quantitative Framework for Financial Management*, Homewood, Ill.. Richard D. Irwin, Inc., 1969.
43. Phillips, C. F., Jr. *The Economics of Regulation.* Homewood, Ill.. Richard D. Irwin, Inc , 1965.
44. Robichek, A. A. and Myers, S. C. "Conceptual Problems in the Use of Risk Adjusted Discount Rates." *Journal of Finance*, Vol. 21, No. 5 (December 1966), pp. 727–30
45. ——— *Optimal Financing Decisions.* Englewood Cliffs, N.J.· Prentice-Hall, Inc., 1965.
46. ——— "Problems in the Theory of Optimal Capital Structure " *Journal of Financial and Quantitative Analysis*, Vol. 1, No. 2 (June 1966), pp 1–35
47. Robinson, P. "Comparable Earnings Revisited." Unpublished Technical Memorandum, Management Sciences Division, American Telephone and Telegraph Co., 1972.
48. Roseman, H. "Measuring the Cost of Equity Capital for Public Utilities." *Public Utility Law*, (1969), pp. 54–67.
49. Ross, L M. "Comments on the Earnings-Price Note." *Stanford Law Review*, Vol 21, No 3 (February 1969), pp. 644–49.
50. Scherer, F. M. *Industrial Market Structure and Market Performance* Chicago: Rand McNally & Co., 1970.
51 Sharpe, W. F "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance*, Vol. 19, No 4 (September 1964), pp. 425–42.
52. Sher, W "The Cost of Capital and Corporation Finance in the Case of Risk." Paper presented at the Econometric Society Meeting, 1968.
53. Solomon, E. "Alternative Rate of Return Concepts and Their Implications for Utility Regulation." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 65–81.
54. ——— "Leverage and the Cost of Capital." *Journal of Finance*, Vol. 18, No. 2 (May 1963), pp. 273–79.

55. Sparrow, F. T "Mathematical Models of Growth-Allowance for Public Utility Regulation: A Synthesis." *Management Science*, Vol. 13, No. 6 (February 1967), pp. B327–41.
56. Stiglitz, J. E "A Re-examination of the Modigliani-Miller Theorem." *The American Economic Review*, Vol 59, No. 5 (December 1969), pp. 784–93.
57. Tobin, J. "Liquidity Preference as Behavior towards Risk " *Review of Economic Studies*, Vol. 25, No. 1 (February 1958), pp. 65–86.
58. U. S. Federal Communications Commission. *American Telephone and Telegraph Co.* Prepared Testimony, M. J Gordon. F.C.C Docket 16258, Staff Exhibit 17, 1966.
59. ——— *American Telephone and Telegraph Co.* Prepared Testimony, S. C. Myers. F.C.C. Docket 19129, 1971
60. ——— *Hearing on the A.T.&T. Interim Revision and Order.* Sec 2 6, 1967.
61. Vickers, D. *The Theory of the Firm. Production, Capital, and Finance.* New York. McGraw-Hill Book Co., 1968.
62 Westfield, F. M. "Regulation and Conspiracy." *The American Economic Review*, Vol. 55, No. 3 (June 1965), pp. 424–43.

An econometric model of the world copper industry

Franklin M. Fisher
Professor of Economics
Massachusetts Institute of Technology

and

Paul H. Cootner
C. O. G. Miller Professor of Finance
Stanford University

in association with

Martin Neil Baily
Assistant Professor of Economics
Massachusetts Institute of Technology

This paper discusses a complete model of the world copper economy. Supply equations for primary copper are estimated for four principal producing countries and the rest of the world; demand equations are estimated for the United States, Europe, Japan, and the Rest-of-World. Scrap supply equations are estimated for the U. S. and non-U. S. sectors as are price adjustment equations. The model is closed with a net input equation for the United States and various identities. The two separate copper markets (U. S. and non-U. S.) are converted by inputs, the generally free London Metal Exchange and scrap markets, and the price-setting behavior of U. S. producers. The copper market is found to be characterized by low short-run but very high long-run price elasticities, making for considerable sensitivity to exogenous forces. The model, fitted to 1948–1968 data, is used for forecasting and simulation experiments. Generally, short-run forecasts are good and longer-run forecasts not very satisfactory. Perhaps the most interesting finding of the simulation

---

Franklin M. Fisher received the Ph.D. in economics from Harvard University (1960), where he was also an undergraduate. He is a consultant and director of Charles River Associates, an economic research firm. Professor Fisher is a fellow of the American Academy of Arts and Sciences, and has published several works in the field of econometrics. He is a fellow of the Econometric Society and the editor of its journal.

Paul H. Cootner received the B.S. and M.S. degrees from the University of Florida (1949 and 1950, respectively) and the Ph.D. in economics from M.I.T. (1953). Dr. Cootner is currently C.O.G. Miller Professor of Finance at Stanford University. His recent research has been in the commodities futures markets and in economic efficiency in education.

Martin Neil Baily received the Ph.D. degree in economics from M.I.T., where he is now Assistant Professor of Economics. He trained previously in economics at Cambridge University and at Simon Fraser University.

The authors are indebted to Nancy Greene, Brendon Horton, and Alan Marin for research and computational assistance and to Paul Rosenstein-Rodan for helpful comments. They also greatly benefited from discussions of the technology, structure, and marketing practices of the copper industry with many persons at the Corporacion del Cobre de Chile. Sonia Klein and David McNicol, also working on copper models, participated in helpful discussions. Errors are the authors' own responsibility, however.

[illegible]
is the prediction that Chilean output would be very sluggish even in the absence of nationalization and that Chilean revenues would be substantially increased were the Chilean government to increase domestic mine production and to allow world prices to adjust accordingly.

1. Introduction

■ In the conventional view, governmental regulation of economic activity arises from imperfections in industrial organization, which lead to less-than-optimal output and a socially undesirable distribution of income. In that model, the role of regulation is to constrain the regulated industry to behave more like its competitive counterparts—to increase output and reduce price below the levels that would otherwise prevail. The classical economic literature of regulation is oriented toward techniques for most efficiently achieving this goal or toward measuring the degree to which actual regulatory efforts are successful.

When we compare this regulatory model with practice, we find that there is another large fraction of governmental activity based on a quite different model. Even within the classical "regulated industries" sector much of the regulatory effort is designed to require a level of output distinctly different from the competitive standard, sometimes a higher level to subsidize certain consumers, or a restricted level to subsidize certain producers or promote certain social goals. The requirement that railroads maintain unprofitable passenger service or that there be a subsidy of local airline service by long-distance traffic is an example of the former. In the latter case we can mention restrictions on competition between different modes of freight transport or on the interest rates paid on deposits by financial institutions. In these cases, the implicit subsidies and taxes are usually justified by the existence of external economies or diseconomies which render the competitive equilibrium inappropriate or wasteful, although these putative imperfections are often asserted rather than demonstrated. This lack of empirical support leaves the regulatory process open to the charge that it is really a mechanism for redistributing income rather than for ameliorating market imperfections.

Once we accept the idea that governmental regulation of an industry is often undertaken with the goal of income redistribution in mind, we need not necessarily confine ourselves to the rather restricted arena of formal "regulated industries" to see the larger picture unfold. The rather clear theoretical case for regulation of natural gas pipelines broadens naturally to previously unregulated gas production companies and in particular to the redistribution of rents in the latter industry. In the case of restriction of farm output and control of agricultural prices the "monopsony" argument has largely fallen away to reveal the key income redistribution question, particularly since corporate farming has come to the fore. In the United States most of this latter type of "regulation" involves the redistribution of income among its citizens, but in other countries much governmental intervention is obviously keyed to an attempt to redistribute wealth from foreigners to local citizens by control of price and output. This process brings us full circle to the development of cartels (this time government sponsored) such as those which in many of our regulated industries served as the original motivation for intervention. We have seen attempts at such "cartels" in petroleum, cocoa, sugar, wheat, and copper, sometimes with, but often without, the cooperation of the consumer. The scope of action for such cartels is limited not only by the tightness of their controls, but also by the basic supply and demand characteristics of an industry in the absence of a cartel.

This paper was motivated by the interest of the Chilean government in determining the behavior of the world copper market as a prerequisite to governmental planning. Since, in Chile, the industry accounts for a large share of total national output, employment, and exports, the Chileans have use for a model of the world copper economy for purposes of long-range planning. Furthermore, Chile potentially could exercise control over its segment of the copper industry to shift the world distribution of its income in its favor.

The study itself was undertaken with no political goal in mind. It was begun while the copper industry in Chile was still nominally in private hands, the hands of the U. S. producers, and was not completed until the seizure of control had been accomplished. But the model was formulated so that any one of the major producing countries could estimate the impact of various national economic policies upon its share of copper income.

What the model does not include is the secondary impact of one country's efforts to control its national output upon the national policies of other producing countries. Obviously, those secondary impacts will depend in part upon the primary result of the policies and the character of any agreements among the producing countries.

2. The general form of the model

■ This study presents an econometric model of the world copper industry, relatively disaggregated to incorporate different supply equations for each of the major producing countries and different demand equations for each of the principal consuming areas. Largely for reasons of data, the model is an annual one and there is no breakdown into types of copper use. Attention in the construction and use of the model centers on the geographic differences in the industry and on the two-price system generally prevailing, wherein much of the copper in the United States moves at the U. S. producer price and most other copper moves at the London Metal Exchange (LME) price (which, in recent years, has often been well above the U. S. producer price). A general simplified outline of the structure of the model is as follows.

We begin with the U. S. market. The U. S. producers set the U. S. producer price to reflect what they believe to be a sustainable (and profitable) long-run level of copper prices, taking into account their own resulting supply decisions. Believing it in their interest to have a relatively stable price, they then take that price as given for the time being and decide on the amount of copper they will supply at that price. Other countries selling at the U. S. producer price do the same. This determines the supply of primary copper in the United States (except for imports).

Consumers of copper, on the other hand, take price as given. Some of those consumers will be able to buy at the U. S. producer price; some of them will not. The latter consumers can import copper at the LME price or can purchase scrap. Total demand for copper is also strongly influenced by indicators of general industrial activity and by the free market price of aluminum, the principal substitute.

The supply of secondary copper in the United States depends in the model on the total amount of primary copper produced in the past, on the ease with which that copper can be gathered as scrap, and on the scrap price, which is very highly correlated with the LME price, as one would expect.[1] The ease of collection is assumed to be measured by the relative amount of previously produced copper collected in the previous year, being low when that was high and when a great deal of the readily available copper has already been collected.

The difference between U. S. total copper consumption and U. S. available supplies (primary, secondary, and imports) is the change in U. S. inventories. If inventories tend to increase for some period of time and the LME price is below the U. S. producer's price, then producers gradually conclude that they set too high a price and that the proper *long-run* price is below the U. S. producer price. On the other hand, if inventories are being depleted over some period of time and the LME price is well above the U. S. producer price, then the producers gradually conclude that the proper *long-run* price is above the U. S. producer price. They then readjust the U. S. producer price in the indicated direction and the process begins again.

The model for the rest of the world operates in a similar fashion except that the LME price to which it responds is a free market price most of the time. Given expectations about the LME price, producers selling at that price make output decisions and consumers buying at that price make consumption decisions (which are, of course, influenced by other variables as well). Scrap supply depends on factors similar to those operating in the United States. Given the total supply of primary from producing countries, total supply of secondary, and net exports to the U. S., the total supply of copper available for consumption outside the U. S. is determined. The difference between this and total copper actually consumed gives the additions to inventories. The LME price responds to the size of inventories relative to total use of copper.

The two markets are linked in three ways. First, as already stated, the U. S. scrap price is highly correlated with the LME price, as both markets are generally free. Second, there is some flow of copper between the United States and the rest of the world. Finally, producers consider a large difference between the LME and the U. S. producer price as a signal that the U. S. market is probably out of long-run equilibrium and that the U. S. producer price should be adjusted.

3. How and why does the two-market system work?

■ The two-market system described above obviously requires further discussion. In particular, when the U. S. producer price is substantially below the LME price, as has often been the case in recent years, what keeps arbitrage from reducing the LME price and depleting U. S. stocks?[2] Further, why do the U. S. producers consider it in their interest to maintain such a situation?

---

[1] Despite the fact that the two prices are very highly correlated, they do not seem to be equal after allowances for transportation and conversion costs. A simple explanation of this is not readily available.

Part of the answer to the first question is that there do exist forces tending to bring the two prices closer together, although these forces are neither fast acting nor strong. When the LME price is above the U. S. price for an extended period because demand is high, U. S. stocks do get drawn down and the producer price, adjusted upward. Nevertheless, a high degree of arbitrage does not take place, presumably because in such a situation, when U. S. producers ration their sales, favored customers are unwilling to jeopardize their long-run relations with the producers for the sake of the short-run gains from arbitrage. Presumably, producers can fairly readily discover if a large customer is reselling on the LME.

The question remains, however, as to why it is in the interests of the producers to act in this way rather than to adjust the price upward immediately to the market-clearing level and to reap the short-run returns from selling to more customers at higher prices. The basic answer seems to lie in the lag structure of behavior on both sides of the copper market. It will be a repeated theme of this study and is well borne out in the results below that speeds of adjustment in the copper industry are very low. On the supply side this comes from the length of time necessary to bring new mines into production and the high investment in the working of existing mines. On the demand side it occurs, at least in part, because of the length of time required to readapt some copper-using equipment to aluminum. Such long lags mean that the price elasticities of demand and supply are both substantially higher in the long than in the short run, and, as we shall see, supply elasticity, in particular, is quite small in the short run and rather substantial in the long run. This leads to a fairly unstable situation in the short run where a small increase in the exogenous factors affecting demand can easily raise short-run market-clearing prices very considerably, even though the price which equilibrates long-run supply and long-run demand may not be much affected.

In this situation, the major U. S. producers may hesitate to raise price to reap short-run rewards for fear of doing two things. First, a large rise might encourage customers who can use aluminum to invest in aluminum-using machinery. Both the experience gained with that machinery and the time required to change back to copper would mean that, when prices fell, such customers would not switch back again for a long time. Indeed, it is not hard to see that, given the fact that the changeover is costly, copper producers may gain more by keeping copper prices to such customers just below the point at which the changeover to aluminum becomes profitable than they would by going above that point and later trying to regain those customers as price fell toward its long-run equilibrium: the regaining of those customers would require a fall in price not merely to its original level but to a level low enough to compensate for the cost of switching back to copper. Since an inability to obtain copper would also provide an incentive to switch to aluminum, this argument suggests that the producers operate by rationing only those customers who cannot switch or for whom the cost of switching is low. This further suggests that manufacturers of wire are generally able to purchase at the U. S. producer price, but it is difficult to obtain direct evidence on this point.[3]

---

[2] Export quotas are part of the answer and enter our model below But they are not the whole story.

A second reason for not raising the U. S. producer price to the short-run market-clearing level may be that such a rise would be taken by other, possibly new, producers as a signal that the long-run price was rising. This would begin to bring new mines into production and, given that the lags are so long, the resulting effects on supply would not disappear when the circumstances substantially raising the short-run market-clearing price ceased to operate. In this situation, the producers may prefer to forego the possible short-run profits rather than encourage new production which will appear on the market when the situation has changed.

Obviously, this second argument assumes that the short-run nature of the given situation is seen differently by the U. S. producers and the operators of potential new mines. However, once having established a system whereby the U. S. producer price is an index of what the U. S. producers believe to be a sustainable long-run price, those producers may hesitate to change it because a new, higher U. S. producer price may be taken to signal a high long-run price and may be acted on as such.

These considerations are difficult to support with hard evidence, although the producers are explicitly disturbed about irreversible substitution effects. They at least make plausible a situation whose existence must be recognized in any adequate model of the copper industry.

4. Distributed lags and estimation methods

■ Since, as stated, many of the crucial reactions in the copper industry take a good deal of time, we have formulated the corresponding equations of the model in terms of distributed lags, with the dependent variable being influenced by past as well as present independent variables. The simplest and best-known formulation of distributed lags, and the one we have employed, is the Koyck or geometric lag,[4] which can also be formulated as a stock adjustment model along the following lines.

We take as an example the supply curve of mine-produced copper in some country. Denote the amount of copper supplied in year t by S_t and the price received by the producers in question by P_t. Given the price, the producers would like to supply an amount S_t^*, which depends on P_t according to the long-run supply equation

$$S_t^* = \alpha + \beta P_t, \tag{1}$$

for example. However, since it takes time to adjust supply, the producers do not immediately move to a new value for S_t^* in response to a new value for price, but only begin to move in that direction. If we assume that it is only possible to move some fixed fraction, μ, of the desired distance in any year, then

---

[3] This implication was first drawn by McNicol [10]

[4] See Koyck [8] and Nerlove [13]

$$S_t - S_{t-1} = \mu(S_t^* - S_{t-1}), \;\; 0 < \mu < 1. \tag{2}$$

Substituting (1) into (2) and rearranging terms:

$$S_t = \mu\alpha + \mu\beta P_t + (1 - \mu)S_{t-1}. \tag{3}$$

Lag (3) and obtain an expression for S_{t-1}, then lag it again and obtain one for S_{t-2}, and so forth. Repeated substitution of the lagged versions into (3) yields

$$S_t = \alpha + \mu\beta \sum_{\theta=0}^{\infty} \lambda^\theta P_{t-\theta}, \;\; \lambda = 1 - \mu, \tag{4}$$

so that supply in year t depends on present and past prices with the weights given to lagged prices declining geometrically with the length of the lag. In this model, the short-run effect of price on supply is given by $\mu\beta$, but the long-run effect is given by β itself. If μ is relatively small, so that adjustments take place fairly slowly, then the long-run effect can be much larger than the short-run effect. We would expect this in the case of copper supply.

There is another model which leads to the form (4) and its equivalent (3) (other than postulating (4) directly). This is to suppose that adjustment can in fact be completely made but that supply depends upon *expected* price, with expectations being formed in an adaptive way, as follows. Suppose that the supply curve is given by

$$S_t = \alpha + \beta P_t^*, \tag{5}$$

where P_t^* denotes expected long-run price. Suppose further that price expectations are formed by revising earlier expectations in the direction of actual prices, according to the relation

$$P_t^* - P_{t-1}^* = \mu(P_t - P_{t-1}^*), \;\; 0 < \mu < 1. \tag{6}$$

Here μ is no longer the speed of adjustment of supply, but rather the speed of adjustment of price expectations. Equation (6) is equivalent to

$$P_t^* = \mu P_t + (1 - \mu)P_{t-1}^*, \tag{7}$$

so that expected price this year is a weighted average of expected price last year and actual price. Repeated lagging of (7) and substitution for lagged values of P_t^* yields

$$P_t^* = \mu \sum_{\theta=0}^{\infty} \lambda^\theta P_{t-\theta}, \;\; \lambda = 1 - \mu, \tag{8}$$

so that expected price is also a weighted sum of present and lagged prices, with the weights declining geometrically with the length of the lag. Substitution of (8) into (5) now yields (4).

Despite the fact that the two models are equivalent as regards the final supply equation, the stock-adjustment version makes much more sense than the adaptive expectations version for relations in the copper industry. It is perfectly clear that adjustments do take a considerable time; moreover, price expectations are not likely to be formed in the way described by (6) to (8), since, especially in the U. S. market, participants are likely to take current price as a much better index of relatively long-run price than a weighted average of past prices, with λ much different from zero. Our interpretation will run in terms of the stock-adjustment model, therefore, and we have made no attempt to incorporate simultaneously the adaptive expectations feature.

Estimation of the stock-adjustment model requires some care. The easiest form in which to estimate it is (3), but merely estimating this by ordinary least squares (aside from the fact that P_t is an endogenous variable) will lead to inconsistency if the error term is autocorrelated. Such autocorrelation is quite likely in such models. It is not possible to treat satisfactorily a very general model of autocorrelation, and we have settled for a model in which the error term in question is first-order autocorrelated, an assumption which probably comes as close to the truth as any which the data will support. Thus, denoting the error term in (3) by u_t, and assuming it enters additively, we assume

$$u_t = \rho u_{t-1} + e_t, \tag{9}$$

where e_t is assumed to have expectation zero and variance-covariance matrix $\sigma^2 I$, thus not itself being autocorrelated. For cases (not the supply equations) in which the right-hand side variables of such equations are predetermined (being either exogenous to the model or lagged endogenous variables), we proceed by choosing estimates of ρ and the other parameters to minimize the sum of squares of e_t. This is a consistent estimator and, if e_t is normally distributed, it is also maximum likelihood. On the normality assumption, asymptotic standard errors can be computed.[5]

When (as in the supply equations) one or more of the righthand side variables are endogenous, an adaptation of the same technique, due to Fair, is used.[6] This is an instrumental variables technique which takes care of simultaneous equations bias as well as autocorrelation of the sort (9). It requires that among the instruments be the current and once lagged values of all predetermined variables in the equation and the lagged values of all endogenous variables in the equation. In the results below, we indicate the instrumental variables used in estimating each equation.

Before closing this section, we may remark that the autocorrelation assumption (9) is restrictive, not merely because it is first-order, but also because it is assumed that, given u_{t-1}, u_t is not correlated with the *past* values of disturbances from other equations in the model. This is far better than assuming it is uncorrelated with the *current* values of such disturbances, but it is still a strong assumption. More general assumptions are very difficult to handle, however.

Estimation of the supply equations for scrap copper presents certain special difficulties, since those equations, as will be seen below, are nonlinear in the parameters. We shall discuss the methods used in the section on the scrap supply equations.

---

[5] See Cooper [2] for details

[6] See [4]. Fair observes that the asymptotic standard errors for his estimator are difficult to calculate; but this is not the case, since instead of replacing sample values by probability limits, one can calculate the sample version of the asymptotic standard errors, relying on the fact that the formulas are only asymptotic in any case.

5. Supply of primary copper

■ In this section, we report the results for estimation of supply curves for primary copper (mine production). Separate supply curves were estimated for the United States, Chile, Zambia, Canada, and the Rest-of-World. The relative importance of the various suppliers can be seen from the figures for 1963, Table 1.[7]

TABLE 1

MINE PRODUCTION OF COPPER, 1963 (THOUSANDS OF METRIC TONS)

UNITED STATES	1100 6
CHILE	601 1
ZAMBIA	588 1
CANADA	410.6
REST–OF–WORLD[a]	1174 4
TOTAL[a]	3874 8

[a] EXCLUDES EASTERN–BLOC COUNTRIES.

The principal suppliers for which separate supply curves were not estimated are Peru, Republic of Congo, and Japan. Data problems prevented such estimation for the Congo; Japanese production did not seem large enough to warrant a separate effort; and an attempt to estimate a separate equation for Peru failed.

It will be noticed that a supply curve for Chile was estimated. Since part of the purpose of the model is to examine the likely effects of alternative Chilean policies, that supply curve is used only for purposes of comparison. Forecasts will be made which examine the effects of departures from the supply relationship which has obtained in the past.

For the United States, the price of copper used in the supply equation was the *Engineering and Mining Journal* (EMJ)[8] price deflated by the U. S. wholesale price index. The EMJ price is a weighted average of the U. S. producer price and the LME price, with the U. S. producer price getting more than 97.5 percent of the weight (this equation is reported in the section on prices). It apparently reflects the prices at which copper is actually traded in the United States a bit more accurately than does the U. S. producer price. Nevertheless, according to the description of the model given in the opening section, it ought to be the U. S. producer price itself which enters the supply relation. Experimentation shows that the results obtained using the EMJ price are uniformly better and more reasonable than those obtained using the U. S. producer price directly, and we have proceeded on that basis. This phenomenon can be rationalized by observing that when the EMJ and U. S. producer prices get substantially out of line, the chances are good that the latter price will be revised so that the EMJ price may capture long-run expectations a bit better than does its principal component.

Denoting U. S. mine production (thousands of metric tons) by USMP and the EMJ price (cents per pound) divided by the U. S. wholesale price index (1957–1959 = 1.00) by USP_{EMJ}, the estimated U. S. supply equation is [9]

$$
\begin{array}{llll}
USMP_t = -160.04 + & 14.27\,USP_{EMJt} + & 0.7261\,USMP_{t-1} & \\
 & (4.769) & (0.2043) & \\
 & (2.996) & (3.554) & (10)
\end{array}
$$

$$\rho = 0.5 \qquad \text{Years: 1949–1958, 1962–1966.}$$

In this, as in all later equations (except where noted), the figures in the first line of parentheses are the asymptotic standard errors of the corresponding estimated coefficients; the figures in the second line of parentheses are the ratios of the estimated coefficients to their asymptotic standard errors. Small sample significance tests are not known for such estimates, but a good rule of thumb is that a coefficient at least twice its asymptotic standard error indicates a statistically significant relationship.

---

[7] "Rest-of-World" in any section of this study denotes the countries not explicitly studied in that section.

[8] Data sources are listed and discussed in Appendix 1.

[9] As with all the equations, the "Years" figures indicate the observations on the dependent variable directly used in the final regression. Two earlier years of data are used in the estimation method.

The figure denoted as ρ is the estimated first-order autocorrelation coefficient of the disturbance in the equation; see equation (9), above. The final estimates are obtained by searching over alternate values of ρ ranging from -1.0 to $+1.0$ by steps of 0.1 and choosing those results (for all parameters) for which the sum of squares being minimized is least.

As indicated, the years 1959–1961 and 1967–1968 have been omitted (our data run through 1968 in other equations). This has been done to eliminate the effect of major copper strikes in the U. S. in 1959 and 1967–1968.[10]

The slow speed of adjustment in U. S. copper supply is indicated by the coefficient of lagged mine production. Only a little more than a quarter (1.00–0.73) of the gap between desired production and actual production is closed each year (see the discussion in Section 4). This is naturally reflected in a fairly large difference between short- and long-run supply elasticities. At the point of means for the period, the short-run price elasticity of supply is approximately 0.453, while the long-run elasticity is approximately 1.67. U. S. adjustment speed, however, while slow, appears faster than the adjustment speed of some of the other producing countries. We now turn to the results for those other countries.

There is some difficulty in deciding on the appropriate price variable to use in the Chilean supply curve because of the effect of the special exchange rates which have been used to tax the copper producers and because of the Chilean inflation. Fortunately, Mamalakis and Reynolds[11] calculate a series for the price received by the producers and, while this only goes through 1959, it has been brought up-to-date by Vittorio Corbo Lioi, who kindly made it available to us. We denote that price (the money price is taken as 1.00 in 1965 and deflated by the Chilean wholesale price index, which is 1.00 in 1958) by ChP. Denoting Chilean mine production (thousands of metric tons) by ChMP, the results are as follows:

$$\begin{array}{llll} \text{ChMP}_t = 91.37 + & 415.4\,\text{ChP}_t + & 0.7206\,\text{ChMP}_{t-1} & \\ & (164.9) & (0.1309) & \\ & (2.520) & (5.505) & \qquad (11) \end{array}$$

$$\rho = -0.1 \qquad \text{Years: 1948–1968.}$$

The speed of adjustment is just about the same as in the United States. At the point of means for the period, the short-run elasticity of supply is approximately 0.112, rather less than for the United States. Long-run elasticity, moreover, is approximately 0.402, considerably less than the comparable U. S. figure. Chilean mine production does not appear to have been very price sensitive.[12]

---

[10] The data began before 1949, but the lags involved in the equation and the estimation method used mean the loss of two years at the start of the time period. Similarly, eliminating 1959 means eliminating 1960 and 1961 from direct use as observations.

[11] See [11].

[12] For comparison, a similar equation using the EMJ price and the U. S. wholesale price index was estimated. It differed from (11) in having a *very* slow speed of

For Canada, we used the EMJ price converted to Canadian dollars and deflated by the Canadian wholesale price index (1958 = 1.00) as the price variable.[13] Denoting this by CanP_{EMJ} and denoting Canadian mine production (thousands of metric tons) by CanMP, the results are:

$$\begin{array}{lllll} \text{CanMP} = -43.73 & + \; 2.129\,\text{CanP}_{\text{EMJ}t} & + \; 0.9873\,\text{CanMP}_{t-1} & & \\ & \quad (0.8738) & \quad (0.03932) & & \\ & \quad (2.437) & \quad (25.11) & & (12) \end{array}$$

$$\rho = -0.4 \qquad \text{Years: } 1948\text{–}1967.$$

Clearly, the speed of adjustment is extremely low, far lower than for the United States or Chile. Less than 2 percent of desired adjustments take place in any one year. At the point of means for the period, the short-run elasticity of Canadian supply is approximately 0.188 and the long-run elasticity is far greater, being estimated at 14.84, although the exact figure is not very reliable.

The situation is similar for the estimated supply curve for Zambia. Here we used the LME price (pounds sterling per long ton) deflated by an index of the cost of living for Europeans in Zambia (1958 = 1.00).[14] Denoting this price by ZP_{LME} and Zambian mine production (thousands of metric tons) by ZMP, the results are:

$$\begin{array}{lllll} \text{ZMP}_t = -69.19 & + \; 0.1269\,\text{ZP}_{\text{LME}t} & + \; 1.103\,\text{ZMP}_{t-1} & & \\ & \quad (0.4446) & \quad (0.3138) & & \\ & \quad (0.2832) & \quad (3.547) & & (13) \end{array}$$

$$\rho = -0.3 \qquad \text{Years: } 1955\text{–}1957,\ 1961\text{–}1965.$$

The poorer quality of these results compared to the others doubtless reflects the poor data and the number of observations that had to be dropped because of strikes and political troubles. At the point of means, the short-run price elasticity of Zambian supply is approximately 0.0684, while the long-run elasticity is *far* greater. Indeed, as estimated, the effects of past prices never die out, although the coefficient of lagged mine production is not significantly above unity;[15] but the speed of adjustment is obviously extremely slow, although *ultimately* the Zambian supply curve is nearly flat. The very high long-run elasticities ought to be expected where new mines are developing and old ones far from exhausted.

---

adjustment and hence a high long-run (but not short-run) price elasticity. The price term was smaller relative to its asymptotic standard error than in (11), however, and we accept (11) as the superior equation. The alternative results were:

$$\begin{array}{lllll} \text{ChMP}_t = -54.43 & + \; 2.740\,\text{USP}_{\text{EMJ}t} & + \; 0.9517\,\text{ChMP}_{t-1} & & \\ & \quad (1.566) & \quad (0.0736) & & \\ & \quad (1.750) & \quad (12.92) & & (11a) \end{array}$$

$$\rho = -0.2 \qquad \text{Years: } 1948\text{–}1968.$$

[13] Some fraction of Canadian supply is sold at the LME price. Preliminary attempts were made to include the LME price in the equation, but they were unsuccessful.

[14] There is not much choice in terms of available data There is also an incomplete series on the cost of living of Africans in Zambia.

[15] In forecasting with the model, this equation is replaced with a trend. See below, p. 597

The remaining supply curve estimated was for the Rest-of-World. The price variable used was the LME price expressed in dollars per long ton deflated by the U. S. wholesale price index (1957–1959 = 1.00).[16] Denoting this by USP_{LME} and Rest-of-World mine production (thousands of metric tons) by RWMP, the results are:

$$\begin{array}{llll} RWMP_t = -28.44 + & 0.2222\ USP_{LME t} + & 0.8832\ RWMP_{t-1} & \\ & (0.09561) & (0.09601) & \\ & (2.324) & (9.199) & (14) \end{array}$$

$$\rho = 0.5 \qquad \text{Years: } 1948\text{–}1968.$$

The speed of adjustment here is still low, but it is higher than in Zambian and Canadian results. At the point of means for the period, the short-run elasticity of Rest-of-World supply is approximately 0.1963 and the long-run elasticity is approximately 1.680.

This completes the results for primary supply. In the estimation of the equations in this section, the instrumental variables used in addition to those required for the Fair method (see the preceding section) were as follows: the lagged ratio of non-U. S. stocks of copper to non-U. S. total use of copper (this appears in the price equations below); the lagged value of whichever of the LME and EMJ price did not appear in the particular equation being estimated; the lagged value of separately accounted for Western Hemisphere total mine production, taken as moving primarily at the EMJ price (United States, Chile, and Canada); and the lagged value of the remaining mine production, taken as moving primarily at the LME price (Zambia and Rest-of-World).[17] In the case of Chile, both the lagged LME and the lagged EMJ price were used as were the two ratios of the Chilean price (ChP) to the other two prices, since those ratios presumably reflect exogenous governmental actions in Chile.[18]

6. Secondary supply

■ We now consider the supply of copper from scrap. Here we divide the world into the United States and the Rest-of-World. The breakdown available in the data is somewhat different for the two regions. Some idea of the order of magnitudes involved may be gained from Table 2, which reports figures for 1963.

TABLE 2

SECONDARY SUPPLY OF COPPER, 1963 (THOUSANDS OF METRIC TONS)

U. S. OLD SCRAP COLLECTION	382.7
U. S. NEW SCRAP	629.3
REST–OF–WORLD TOTAL SCRAP[a]	1328.9
TOTAL[a]	2340.9

[a] EXCLUDES EASTERN–BLOC COUNTRIES.

Old scrap is scrap created by the destruction of old fabricated products; new scrap, on the other hand, is scrap created in the fabricating process itself. It is either returned to the refiner or used again by the fabricator. While it is clear that both types of scrap are part of secondary supply, it is not quite so clear whether new scrap should also be included in the variable representing the demand for copper or whether it should be subtracted. While it is possible to argue on both sides of that question, the results for the demand equations (reported in the next section) are far better when direct use of scrap is included in demand, so *all consumption figures in this study include such scrap.*[19]

---

[16] This is not a very satisfactory deflator, but it is difficult to see how to improve it much.

[17] In the case of the Zambian supply curve (13), where there were relatively few observations, only the sum of the last two variables was used.

[18] In a few cases in which data on instrumental variables were missing for a year or so at the beginning of the period, we extrapolated backwards to construct them. Note that this was not done for the variables actually appearing in the equation being estimated

Obviously, the two different ways in which scrap is generated suggest two different models, and for the United States, the data permit their construction. We begin with the more complex case of old scrap.

The prime determinants of old scrap collections in our model are (1) the amount of copper available for collection in the broadest sense, (2) the ease with which that copper can be collected, and (3) the price paid for it.

In the broadest sense, the amount of copper available for old scrap collection in the United States is the total amount of copper embodied in already produced copper products. The change in the available scrap supply during year t is given by the identity:

Change in available scrap supply = primary production
+ net imports of refined copper
+ net imports of fabricated copper (15)
− increases in stocks of copper.

If we knew the amount of copper available for collection in any given base year, then the amount available in any other year could be calculated from (15). We do not have such a benchmark figure, however, and, as a result, we have estimated that figure as a parameter in our old scrap equation. Thus letting K_t be the copper available at the beginning of year t, we write

$$K_t \equiv K_{1948} + K_t^*, \tag{16}$$

where K_t^* is obtained by cumulating (15) from the beginning of 1948 and K_{1948} is to be treated as a parameter (all figures below are in thousands of metric tons).[20]

It is only in the broadest sense, of course, that all copper in the United States is available for collection. In part, we account for this in our model by explaining what fraction of K_t is in fact collected. It must be recognized, however, that K_{1948} does not represent the cumulation of (15) from time immemorial, but rather represents the amount of copper in a narrower sense available for scrap collection at the beginning of 1948. This narrower sense applies, of course, to copper produced or imported since 1948, so that K_t^* also is broader than is strictly appropriate. Fortunately, it is easy to see that the fact that not all produced or imported copper is really available for collection matters very little to our model or results.

Denote United States old scrap collections (thousands of metric tons) in year t by USOS_t.[21] The dependent variable in our equation below (its lagged value will also appear) will be $\log(\text{USOS}_t/K_t)$. Now suppose that instead of K_t being available, some constant fraction of K_t, say δ, is available. Then the true dependent variable should be $\log(\text{USOS}_t/\delta K_t) = \log(\text{USOS}_t/K_t) - \log \delta$. Hence the only effect of the difference between using K_t and δK_t is to place a term in $\log \delta$ in the constant term of our results.

---

[19] This follows usage in the data sources which call such a variable "total use." Note that "consumption" in the sources does not include direct use of scrap as do our "consumption" variables. (This is not the same issue as that discussed in the text.) See Appendix 1 for the way in which new scrap figures were calculated and for further discussion.

[20] 1948 is used as a benchmark because changes in copper stocks are unknown before that date. The estimated equation begins with 1950 because K_{t-2} is required by the estimation method. A similar statement holds for the Rest-of-World, p. 583, below, where K_{1950} is taken as a benchmark and estimation starts with 1952.

Obviously, this argument makes it convenient to use a logarithmic model here; moreover, it is natural to do so, since it is natural to think in terms of explaining the fraction of available scrap which is actually collected.

As a proxy for a measure of the difficulty of collection, we use the fraction of available scrap that was collected as such in the preceding year. This is not a very direct measure of difficulty, but it makes sense if we think of available copper as in forms and locations of differing ease of collection. One expects the copper easiest to collect to be collected first; hence if a relatively large amount of the available copper was collected last year, copper is likely to be relatively hard to collect this year. Accordingly (and quite unusually) we should expect the coefficient of the lagged dependent variable to be *negative.*

There is a separate scrap price in the United States, but it is very highly correlated with the LME price, as one would expect, since both are free market prices.[22] We found that it made little difference to the results which price was used, and, accordingly, we simplified the model slightly by using the LME price expressed in dollars per long ton and deflated by the U. S. wholesale price index (1957–1959 = 1.00). We denote that price by USP_{LME}.

Unfortunately, the estimation of the old scrap supply equation runs into a minor but rather annoying difficulty. The equation is not linear in K_{1948}. Accordingly, we proceeded by choosing alternate values for K_{1948} and then estimating the resulting equation, intending to choose the version in which the final sum of squares was minimized.[23] As it turns out, however, the results are quite insensitive to the value of K_{1948} chosen, suggesting that any estimate thereof will have a very large asymptotic standard error. (The asymptotic standard errors reported below are all conditional on the values of K_{1948} chosen.) This would not be surprising—although it does not occur in the results for the Rest-of-World, reported below—since it indicates that scrap collections are not very dependent on copper produced before 1948; however, the sum of squares to be minimized continues to decline slowly as a function of K_{1948} until that parameter is well beyond any reasonable value.

---

[21] The figure actually used was collection of old scrap. This includes a small amount of *imported* scrap which does not precisely fit the model being discussed.

[22] As already mentioned, it does not appear to be the case that the two prices are equal after allowance for transportation and conversion costs We are unable to explain this.

[23] The actual estimation procedure used was fairly complicated. First, USP_{LME} was regressed on its own lagged value; the lagged value of the EMJ price; the lagged change in the ratio of stocks outside the United States to total consumption outside the United States (see Section 8); the lagged sum of mine production in the U S , Chile, and Canada; and the similar lagged sum for all other countries. This yielded a predicted value of USP_{LME}, which we shall call $\widehat{\text{USP}}_{\text{LME}}$. Then, for each choice of

Fortunately, this makes essentially no difference to our estimate of the price sensitivity of old scrap supply, our estimate of the effect of former collections, or our implied forecast of old scrap collections. We therefore present results for high and for low values of K_{1948} as well as for a value chosen to bear roughly the same relation to U. S. consumption as the similarly estimated figure for the Rest-of-World (below) bears to Rest-of-World consumption. That value of K_{1948} is 60,000 thousand metric tons. When it is used, we obtain

$$\log\left(\frac{\mathrm{USOS}_t}{60{,}000 + \mathrm{K}_t^*}\right) = -9.878 - \underset{\substack{(0.1261)\\(-2.960)}}{0.3731} \log\left(\frac{\mathrm{USOS}_{t-1}}{60{,}000 + \mathrm{K}_{t-1}^*}\right) \tag{17}$$

$$+ \underset{\substack{(0.1064)\\(3.968)}}{0.4222} \log \mathrm{USP}_{\mathrm{LME}t}$$

$$\rho = 0.9 \qquad \text{Years: 1950–1968}$$
$$\text{Sum of Squared Residuals} = 0.08338.$$

By way of comparison, if we use alternate estimates of K_{1948} of 20,000 and 140,000 metric tons, we obtain, respectively,[24]

$$\log\left(\frac{\mathrm{USOS}_t}{20{,}000 + K_t^*}\right) = -9.093 - 0.3761 \log\left(\frac{\mathrm{USOS}_{t-1}}{20{,}000 + K_{t-1}^*}\right) \tag{18}$$

$$+0.4371 \log \mathrm{USP}_{\mathrm{LME}t}$$

$$\rho = 0.9 \qquad \text{Sum of Squared Residuals} = 0.08755$$

and

$$\log\left(\frac{\mathrm{USOS}_t}{140{,}000 + K_t^*}\right) = -10.88 - 0.3822 \log\left(\frac{\mathrm{USOS}_{t-1}}{140{,}000 + K_{t-1}^*}\right)$$

$$+ 0.4417 \log \mathrm{USP}_{\mathrm{LME}t} \tag{19}$$

$$\rho = 0.7 \qquad \text{Sum of Squared Residuals} = 0.07208.$$

In all these equations, the short-run elasticity of old scrap supply with respect to the LME price is about +0.42 to +0.44. The implied long-run elasticity is lower because high scrap collections in one period mean lower ones in the next period (other things equal), being about +0.31 to +0.32. The different choices for K_{1948} affect only the second decimal place, and that only in a very minor way.

Our model for new scrap is much simpler; we estimated it as a linear function of total consumption. Denoting U. S. new scrap

---

K_{1948}, Fair's method [4] was used to estimate the final equation, with the instrumental variables being those required by the method plus the logarithm of $\widehat{\mathrm{USP}}_{\mathrm{LME}}$. Note that different Fair method instruments arise for different choices of K_{1948} and that each such estimation requires a search over values of ρ.

[24] Asymptotic standard errors are not presented in (18) and (19) since they are for comparative purposes only and the extra computation does not seem warranted. They would be much the same as in (17).

collections by USNS[25] and total consumption (total use) by USC (both in thousands of metric tons), the estimated equation is[26]

$$
\begin{aligned}
\mathrm{USNS}_t = -275.2 &+ 0.3961\ \mathrm{USC}_t \\
&\quad (0.0524) \\
&\quad (7.555)
\end{aligned}
\qquad (20)
$$

$$\rho = 0.2 \qquad \text{Years: 1947–1968.}$$

We could find no evidence of a significant price effect here. The elasticity with respect to USC is 1.48 at the point of means.

Outside the United States, the data do not permit a breakdown into old and new scrap and we estimated a single equation for total scrap supply. This was necessarily a hybrid of the two types estimated for the United States. Denoting Rest-of-World secondary supply by RWS and Rest-of-World total consumption by RWC* (both in thousands of metric tons), the results are as follows:[27]

$$
\begin{aligned}
\log\left(\frac{\mathrm{RWS}_t}{53{,}000 + K_t^*}\right) = -5.241 &- \underset{\substack{(0.4729)\\(-1.328)}}{0.6278}\log\left(\frac{\mathrm{RWS}_{t-1}}{53{,}000 + K_{t-1}^*}\right) \\
&+ \underset{\substack{(0.09879)\\(2.557)}}{0.2546}\log \mathrm{USP}_{\mathrm{LME}t} + \underset{\substack{(0.3204)\\(2.978)}}{0.9534}\log\left(\frac{\mathrm{RWC}_t^*}{53{,}000 + K_t^*}\right)
\end{aligned}
\qquad (21)
$$

$$\rho = 0.2 \qquad \text{Years: 1952–1968.}$$

Short-run price elasticity is about +0.25 and short-run elasticity with respect to total copper consumption about +0.95. The corresponding long-run figures are about +0.16 and +0.52. In comparing these to the U. S. figures (recalling the hybrid nature of the equation), it is clear that they are not very different.

7. Demand

■ We come now to the estimation of demand equations for copper.[28] Separate demand equations were estimated for the United States, Europe, Japan, and the Rest-of-World. Attempts to estimate separate equations for individual European countries resulted in unsatisfactory demand equations, although the results for Europe as a whole were reasonably satisfactory. Why this should be true is hard to say. Aside from purely statistical reasons, such as the possibility that the disturbances in the demand equations in two neighboring countries may be negatively correlated, a partial explanation may lie in either of two facts. First, individual European countries may differ in the skill with which they manage to buy on the LME. If the success of their buying agents varies over time, the use of an annual LME figure may be better for Europe as a whole than for individual countries. Second, we have treated industrial production as a single aggregate. If intra-European exports have a different copper-using component than European industrial production generally, then an aggregate relation in which such exports are netted out may do better than individual ones. Neither of these explanations seems wholly adequate, however.[29]

---

[25] Our new scrap figures, for reasons of consistency, were taken as direct use of scrap plus secondary refined less old scrap. See Appendix 1, Table 9.

[26] Estimation was by Fair's method with additional instrumental variables, lagged EMJ and LME prices, lagged U. S. mine production, and U. S. industrial production.

[27] The estimation method was similar to that described above for U. S old scrap, except that we formed not only $\widehat{\mathrm{USP}}_{\mathrm{LME}}$ and used its logarithm as an instrumental variable, but also $\widehat{\mathrm{RWC}}{}^*$ and used *its* logarithm as an instrumental variable. In this first stage of the procedure, the same regression was used as before to form $\widehat{\mathrm{USP}}_{\mathrm{LME}}$. In the formation of $\widehat{\mathrm{RWC}}{}^*$, the regressors were the lagged EMJ and LME prices, lagged RWC*, and the two lagged sums of mine productions.

[28] As used in this study, "demand" means "total use." See note 19 above.

TABLE 3

TOTAL USE OF COPPER, 1963 (THOUSANDS OF METRIC TONS)

UNITED STATES	2320.4
EUROPE	2631.8
JAPAN	592.1
REST-OF-WORLD[a]	526.0
TOTAL[a]	6070.3

[a]EXCLUDES EASTERN-BLOC COUNTRIES

The relative importance of the various areas in copper consumption is shown by the figures for 1963 in Table 3. We begin with the United States, for which data are best. It is clear that the principal elements of a copper demand equation must be the price of copper, the price of substitutes, and measures of industrial activity. For the first, we used USP_{EMJ}, the EMJ price deflated by the U. S. wholesale price index. The remaining variables require some discussion, however.

The principal substitute for copper is, of course, aluminum. The obvious choice for a price to use is the deflated U. S. price of aluminum (deflation takes care of more general substitutes for copper). When this was attempted, the attempt failed, but it failed in an informative way. Without the inclusion of any aluminum price, a reasonably satisfactory looking demand equation was obtained. When the U. S. aluminum price was included, however, not only did its coefficient fail to be positive (as should be the case for a substitute), but it was significantly negative and the whole equation changed radically. One cannot properly conclude from this merely that aluminum substitution fails to show up in the U. S. demand equation for copper; it does show up in a way which makes no economic sense but which cannot be ignored.

Fortunately, the reason for this is fairly easy to find. The U. S. producer price of aluminum is an administered price which, like the producer price of copper, is changed only infrequently. Unlike the copper price, however, the aluminum price is often an unrealistic quotation which is either widely discounted or produces severe rationing on other occasions.[30] A reasonably satisfactory way to take account of this artificial, administered price is to use a free market aluminum price as an index of the real cost of obtaining aluminum. Since the London price was also controlled for part of the period, the most readily available price is the German price of aluminum in deutsche mark per metric ton and we used this converted to dollars and deflated by the U. S. wholesale price index (1957–1959 = 1.00). It is denoted by USAlP, but the initials represent the deflator and not the source of the money price.[31]

---

[29] It is true that our estimates ignore the two episodes of attempted interference with the LME in the middle 1950s and middle 1960s. This may have affected the results (although it is hard to see just how) but has little to do with the aggregation question discussed in the text.

[30] In the Fall of 1971, for example, one producer lowered the price from $.29 to $.23, stating that the reduction merely reflected "ridiculous" discounting.

Since even a short-run adjustment to prices is likely to be delayed, we used both copper and aluminum prices lagged by one year, so that demand depends on last year's (and previous years') prices.

There is more than one possible choice for an index of copper-using industrial activity. The two most obvious are the Federal Reserve Board index of industrial production and index of construction materials.[32] Ideally, one would want to use both, but they are so correlated that no sensible result is obtained when both are included. We present results below using each. The index of industrial production (1963 = 100) is denoted by USIP and the index of construction materials (1957–1959 = 100) by USCM.

One more matter needs to be discussed before proceeding to the results. A demand equation should not include changes in inventories of copper. Changes in inventories of copper *held as such* are not included in the consumption data; they are treated separately below. There is another way in which copper inventories can be held, however; that is in the form of fabricated products. Unfortunately, there are no data on the inventories of copper fabricated products. There are data, however, on inventories of durable goods. If we assume that the change in the copper content of such inventories is reasonably linearly related to the total change in the inventories themselves, then that total change should be included in the demand equation. We denote it as ΔUSID. (It is measured in billions of dollars deflated by the U. S. wholesale price index, 1957–1959 = 1.00.)

This argument has an important further and testable consequence. We are using a Koyck distributed lag in copper demand, as described in Section 4, above, and already used for supply. The logic of the adjustment model leading to that lag suggests that the appropriate lagged variable to use is not lagged copper consumption as measured, but lagged copper consumption corrected for the (lagged) change in durable inventories. This, however, leads to the following.

The basic assumption on which we include the change in inventories in the equation is that the change in copper inventories held as finished goods is linearly related to the total change in inventories; that is, that the change in such copper inventories in year t is well approximated by a term $(\alpha + \beta\Delta \text{USID}_t)$. Denoting U. S. copper consumption in year t as measured (thousands of metric tons) by USC_t,[33] the stock adjustment model implies that instead of having USC_t on the left and λUSC_t on the right (where $\mu = 1 - \lambda$ is the speed of adjustment; see Section 4), we should have $(\text{USC}_t - \alpha - \beta\Delta\text{USID}_t)$ on the left and $\lambda(\text{USC}_{t-1} - \alpha - \beta\Delta\text{USID}_{t-1})$ on the right; so that (letting γX_t stand symbolically for all the other variables already discussed) the full demand equation reads

---

[31] We have treated the aluminum price as exogenous to the copper market (in the short run) which is somewhat questionable but not too poor an approximation.

[32] See Federal Reserve System [5].

[33] This includes copper that will become new scrap. See above, pp. 579–80.

$$\begin{aligned}\text{USC}_t - \alpha - \beta\Delta\text{USID}_t = \delta + \gamma X_t & \\ + \lambda(\text{USC}_{t-1} - \alpha - \beta\Delta\text{USID}_{t-1}), & \quad (22)\end{aligned}$$

where δ is a parameter. Rearranging this, we obtain

$$\begin{aligned}\text{USC}_t = (\delta + \alpha - \lambda\alpha) + \gamma X_t & \\ + \beta\Delta\text{USID}_t - \lambda\beta\Delta\text{USID}_{t-1} + \lambda\text{USC}_{t-1}, & \quad (23)\end{aligned}$$

so that not only should ΔUSID_{t-1} be included, but its coefficient should turn out to be minus the product of the coefficients of ΔUSID_t and USC_{t-1}. It would, of course, be possible (although slightly cumbersome) to impose this constraint in estimating the equation, but it is far preferable not to impose it and to see whether it is approximately satisfied in the results. If it is, we have gained a check on a joint implication of the adjustment model assumed and on the argument as to inventories of copper held in finished form. In this event, the constraint is indeed approximately satisfied, thus reinforcing our faith in the assumptions made.

We are now ready for the results. When the index of construction materials is used, the estimated demand equation is[34]

$$\begin{aligned}\text{USC}_t = {} & -194.7 - \underset{\substack{(1.903)\\(-8.020)}}{15.26}\,\text{USP}_{\text{EMJ}t-1} + \underset{\substack{(4.794)\\(2.028)}}{9.721}\,\text{USAIP}_{t-1} \\ & + \underset{\substack{(1.265)\\(5.557)}}{7.029}\,\text{USCM}_t + \underset{\substack{(5.904)\\(9.166)}}{54.11}\,\Delta\text{USID}_t \qquad (24) \\ & - \underset{\substack{(7.667)\\(-5.213)}}{39.97}\,\Delta\text{USID}_{t-1} + \underset{\substack{(0.1218)\\(6.046)}}{0.7363}\,\text{USC}_{t-1}\end{aligned}$$

$R^2 = 0.991 \qquad \rho = -0.8 \qquad$ Years: 1950–58, 1962–1966.

When the index of industrial production is used, the results are similar:

$$\begin{aligned}\text{USC}_t = {} & -14.75 - \underset{\substack{(1.752)\\(-7.060)}}{12.37}\,\text{USP}_{\text{EMJ}t-1} + \underset{\substack{(4.642)\\(1.786)}}{8.290}\,\text{USAIP}_{t-1} \\ & + \underset{\substack{(0.9134)\\(5.559)}}{5.078}\,\text{USIP}_t + \underset{\substack{(6.413)\\(9.431)}}{60.49}\,\Delta\text{USID}_t \qquad (25) \\ & - \underset{\substack{(7.102)\\(-6.251)}}{44.40}\,\Delta\text{USID}_{t-1} + \underset{\substack{(0.1126)\\(7.024)}}{0.7910}\,\text{USC}_{t-1}\end{aligned}$$

$R^2 = 0.991 \qquad \rho = -0.8 \qquad$ Years: 1950–1958, 1962–1966.

In either case, the results are spectacularly good. Most of the estimated coefficients are several times their asymptotic standard errors, the only exception being the coefficient of aluminum price in (25). It is particularly noteworthy that this is true of the coefficient of copper price, an unusual feature for econometric estimates of demand equations. More important, every coefficient has the expected sign and the constraint on the coefficient of $\Delta USID_{t-1}$ is very closely satisfied. In (24), that coefficient is -39.97, whereas the product of the coefficients of $\Delta USID_t$ and USC_{t-1} is 39.84; in (25), the coefficient is -44.40, whereas the corresponding product is 47.84.

---

[34] In equations estimated by the Hildreth-Lu technique, $1 - R^2$ is the sum of squares of errors in the original equation to be estimated divided by the centered sum of squares of the dependent variable in that equation.

In terms of elasticities, the two equations are very close. From (24), at the point of means for the period, the elasticity of copper consumption[35] with respect to copper price is -0.2131 in the short run and -0.9002 in the long run. From (25), the corresponding figures are -0.1727 and -0.8168, respectively. The elasticity with respect to the price of aluminum implied by (24) is $+0.2392$ in the short run and $+1.010$ in the long run; the corresponding figures from (25) are $+0.2040$ and $+0.9759$, respectively. Finally, the elasticity with respect to the index of construction materials is 0.3318 in the short run and 1.402 in the long run, from (24). The comparable elasticities from (25) with respect to the index of industrial production are 0.1529 and 0.7317, respectively. Note that the last set of elasticities measures roughly the same thing in both equations, since in either equation the variable in question is serving as a proxy for economic activity in general.[36]

When we turn to consuming areas outside the United States, the results are less striking. In large part, this is probably due to the better quality of U. S. data. In particular, much better data on changes in stocks of copper exist for the United States than for most other countries; moreover, only for the United States was it possible to utilize data on stocks of durable goods in general to perform the correction for copper held in the form of finished goods which proved so spectacularly successful in (24) and (25).

The demand equation for Europe is similar to that for the United States, in that it includes copper price, aluminum price, lagged copper consumption, and an index of industrial activity. The copper money price used was, of course, the LME price (pounds sterling per long ton), and the aluminum money price used was again the German aluminum price (deutsche mark per 100 kilograms). Both of these were then converted to dollars. Construction of an appropriate deflator requires some discussion, however.

For each European country, we formed a dollar-equivalent wholesale price index, by taking the country's own wholesale price index (1958 = 1.00) and dividing it by an index of the country's exchange rate vis-à-vis the dollar. The latter was an index of local currency per dollar, scaled to be 1.00 in 1958. These individual country indexes, now in a common unit, were combined in a

---

[35] These are elasticities of USC_t with respect to the several variables, not of copper consumption after the implicit correction for the change in durable goods inventories. The latter cannot be computed without some assumption about the constant term, α, which appears in (22).

[36] The linear form used for the demand equation must be considered only an approximation, especially where such activity variables are concerned, so these last elasticities have perhaps less meaning than do the elasticities with respect to copper and aluminum prices.

TABLE 4

WEIGHTS OF INDIVIDUAL COUNTRIES IN EUROPEAN PRICE INDEX (PERCENT)

GERMANY	26.6
UNITED KINGDOM	21.3
FRANCE	19.6
ITALY	10.0
BELGIUM AND LUXEMBOURG	3.5
SWEDEN	3.5
SWITZERLAND	3.5
NETHERLANDS	3.3
SPAIN	3.3
AUSTRIA	2.0
FINLAND	1.3
PORTUGAL	0.8
GREECE	0.5
IRELAND	0.5
TOTAL	100.0[a]

[a] INDIVIDUAL WEIGHTS AS GIVEN DO NOT SUM TO 100.0 BECAUSE OF ROUNDING.

weighted average, the weights being proportional to 1963 industrial production.[37] The weights are shown in Table 4. Three different indexes of production were tried, all with very similar results. They were the OECD Europe industrial production index,[38] the United Nations index of manufacturing, and the United Nations index of industrial production. The United Nations indexes[39] have somewhat wider coverage than the OECD index. We report the result using the former index of industrial production.

The dependent variable was total use of copper (thousands of metric tons) for all noncommunist Europe. We denote it by EURC. The two prices are denoted by $\mathrm{EURP_{LME}}$ and EURAlP, respectively, and the index of industrial production (1958 = 100) by EURIP.

The estimated European demand equation is

$$\begin{aligned} \mathrm{EURC}_t = {} & -1220 - \underset{\substack{(0.1961)\\(-1.373)}}{0.2693}\,\mathrm{EURP}_{\mathrm{LME}t-1} + \underset{\substack{(24.92)\\(1.144)}}{28.52}\,\mathrm{EURAlP}_{t-1} \\ & + \underset{\substack{(5.795)\\(1.561)}}{9.045}\,\mathrm{EURIP}_t + \underset{\substack{(0.3395)\\(1.598)}}{0.5426}\,\mathrm{EURC}_{t-1} \qquad (26) \\ & R^2 = 0.934 \qquad \rho = -0.1 \qquad \text{Years: 1952–1968.} \end{aligned}$$

The elasticities at the point of means are as follows: with respect to copper price, −0.0878 in the short run and −0.1920 in the long run; with respect to aluminum price, +0.6133 in the short run and 1.341 in the long run; and with respect to industrial production, +0.4534 in the short run and +0.9913 in the long run.

The results indicate a somewhat slower speed of adjustment than was found for the United States. Elasticities with respect to copper price are lower and with respect to aluminum price higher than for the United States. Elasticities with respect to industrial production are roughly the same. It must be remembered, however, that the results for Europe are not nearly so reliable as those for the United States. Moreover, it is not possible to make all the same adjustments for changes in copper stocks for both the United States and Europe, so that, in particular, the effect of changes in the stocks of copper held as finished goods is included rather than separately accounted for. Given this, it is nontrivial to have obtained a demand equation with all coefficients being of the expected sign and reasonable magnitude and bigger than their respective asymptotic standard errors.

The results for Japanese demand are rather different. Only here in the model could we find no evidence of a lagged adjustment process. This may be due to the very rapid growth of the Japanese economy, since a situation in which a significant fraction of industrial capacity is new each year is also likely to be one in which a significant part of copper demand does not depend on the relative adjustment of old capacity to new price situations. Nevertheless, it seems doubtful that adjustment is so rapid as to be complete in a single year (although it must be remembered the price entered in the model is last year's price). Also related to the rapid growth of Japan is the fact that we would expect the variables with large effect on Japanese copper demand to be those which are related to industrial production. Despite this, we did find a negative effect of copper price; aluminum price did not appear to play any role.

---

[37] These are the weights given by the Organization for Economic Cooperation and Development to its European members in its industrial production index. See Organization for Economic Cooperation and Development, *General Statistics* (Paris. OECD, 1970).

[38] *Ibid.*

[39] See [15]

We present two equations which differ only in the measure of copper-using activity. (The two measures are too collinear to use in the same equation.) The first equation is an index of total industrial production (1955 = 100) denoted by JIP. The second is an index of the production of construction materials (1955 = 100) denoted by JCM.[40] The price variable used was the LME price expressed in yen per long ton, deflated by the Japanese wholesale price index (1958 = 1.00); it is denoted by JP_{LME}. Japanese consumption of copper (thousands of metric tons) is denoted by JC.

The two equations are as follows:

$$\begin{array}{llll} JC_t = 124.2 & -\ 0.0002334\ JP_{LME\,t-1} & +\ 1.723\ JIP_t & \\ & \quad(0.0001312) & \quad(0.0716) & \\ & \quad(-1.780) & \quad(24.06) & (27) \end{array}$$

$$R^2 = 0.982 \qquad \rho = 0.0 \qquad \text{Years: } 1951\text{–}1968$$

and

$$\begin{array}{llll} JC_t = 66.96 & -\ 0.0002316\ JP_{LME\,t-1} & +\ 2.433\ JCM_t & \\ & \quad(0.0001570) & \quad(0.1413) & \\ & \quad(-1.475) & \quad(17.22) & (28) \end{array}$$

$$R^2 = 0.975 \qquad \rho = 0.1 \qquad \text{Years: } 1954\text{–}1968.$$

The results using the index of industrial production seem somewhat more reliable than those using the index of construction materials, although this may merely reflect the fact that they are based on three more years of data. In any case, the implications of both equations are quite similar. There is no difference between short- and long-run elasticities. From (27), the elasticity of Japanese consumption with respect to price is −0.09428 at the point of means for the period; from (28), that elasticity is −0.1184. From (27), at the point of means, the elasticity with respect to the index of industrial production is 0.6014; from (28), the comparable elasticity with respect to the index of construction materials is 0.9921.

Finally, a demand equation was estimated for the Rest-of-World. Here the prices were the LME price and the German aluminum price, both expressed in dollars and both deflated by the U. S. wholesale price index (1957–1959 = 1.00).[41] The former price is denoted, as before, by USP_{LME} and the latter, as before, by USAlP. For a production index, we used that compiled by the United Nations.[42] The production index is denoted by RWIP (1963 = 100). The dependent variable is Rest-of-World total use of copper

---

[40] *Monthly Statistics of Japan.*

[41] This is not a very satisfactory deflator, but it is difficult to see how to improve it significantly.

[42] Excluding the United States, Europe, and Japan. See [15].

(thousands of metric tons) and is denoted by RWC. The results are as follows:

$$
\begin{array}{lll}
\text{RWC}_t = -11.82 - 0.1212\,\text{USP}_{\text{LME}t-1} & + 0.8828\,\text{USAlP}_{t-1} & \\
\qquad\qquad\quad (0.03746) & \quad (3.014) & \\
\qquad\qquad\quad (-3.237) & \quad (0.2929) & \\
\qquad\qquad\qquad + 1.971\,\text{RWIP}_t & + 0.7646\,\text{RWC}_{t-1} & \\
\qquad\qquad\qquad\quad (0.8127) & \quad (0.1613) & (29) \\
\qquad\qquad\qquad\quad (2.426) & \quad (4.740) & \\
R^2 = 0.939 \qquad \rho = -0.3 & \text{Years: 1951–1968.} &
\end{array}
$$

The result as to the effects of aluminum price is unreliable; the remaining coefficients are considerably larger than their asymptotic standard errors. At the point of means, the estimated elasticities are as follows: with respect to copper price, −0.2177 in the short run and −0.9248 in the long run; with respect to aluminum price, +0.1074 in the short run and +0.4561 in the long run; and with respect to industrial production, +0.4087 in the short run and +1.736 in the long run. The price elasticities are fairly similar to those found for the United States.

In general, the demand results show copper demand to be rather inelastic with respect to copper price, even in the long run. The most important determinants of copper demand, as one would expect, are the levels of industrial activity in the consuming countries.

8. Prices

■ So far, there are two prices directly used in the model, the LME price and the EMJ price. As already mentioned, however, the latter is a constructed price which is very closely associated with the U. S. producer price. Indeed, expressing all three in common units, and denoting the three prices by P_{EMJ}, P_{LME}, and P_{Prod}, least squares regression (with no constant term) reveals

$$
\begin{array}{lll}
\text{P}_{\text{EMJ}t} = 0.9762\,\text{P}_{\text{Prod}t} & + 0.01538\,\text{P}_{\text{LME}t} & \\
\qquad\quad (0.00842) & \quad (0.007045) & \\
\qquad\quad (115.9) & \quad (2.183) & (30) \\
R^2 = 0.999 & \text{Years: 1946–1968;} &
\end{array}
$$

the sum of the coefficients is not very far from unity.[43] We have already outlined in Section 2 the way in which the model describes the setting of the U. S. producer price. Essentially, that price is set as an intended maintainable price, being adjusted relatively slowly. Adjustments in it are made in response to indications that the U. S. market is drifting out of equilibrium. One such indication is clearly the accumulation of private stocks of copper in the United

---

[43] While this is true, some minor anomalies that could be produced because the sum is not unity were not recognized until we formulated the forecasts later in this paper. We could (and did) get LME prices far in excess of the producer price and we still had an EMJ price slightly below the producer price, particularly because the weight on the LME price is so small; i e , about the same size as the discrepancy between the sum of the coefficients and one These anomalies have only minor effects, however It should be noted that casual discussions of the weights in the trade suggest a 10-percent weight on LME price.

States; another is a large difference between the U. S. producer price and the LME price. We find that both of these have a definite effect.

Denoting the change in U. S. private stocks of copper during year t by ΔUSS_t and measuring it in thousands of metric tons, we divide it by U. S. consumption in year t to obtain a measure of the relative size of stocks. Assuming that decisions are taken for this year with an eye to last year's variables, we enter the resulting ratio lagged as well as the lagged difference between the LME and the U. S. producer price. Denoting deflation by the U. S. wholesale price index (1957–1959 = 1.00) by the prefix US, as before, the estimated equation (where all prices are in deflated dollars per long ton) is as follows:[44]

$$USP_{Prod\,t} = 320.8 - \underset{\substack{(825.4)\\(-2.248)}}{1856}\ \frac{\Delta USS_{t-1}}{USC_{t-1}}$$

$$+ \underset{\substack{(0.1461)\\(1.840)}}{0.2689}\,(USP_{LME\,t-1} - USP_{Prod\,t-1}) + \underset{\substack{(0.2452)\\(2.438)}}{0.5980}\,USP_{Prod\,t-1} \qquad (31)$$

$$R^2 = 0.556 \qquad \rho = 0.2 \qquad \text{Years: 1952–1966.}$$

Note that while the speed of adjustment is well below unity, it is high relative to that found for supply adjustments and for U. S. demand adjustment. About 40 percent of the gap between desired and actual price is covered in a year. The other two coefficients have the expected sign.[45]

We turn now to the equation explaining the LME price. Here we also expect the size of copper stocks to play the chief role (although, of course, these will be stocks outside the United States); however, the rationale behind this is a bit different from that in the case of the U. S. producer price. Whereas that price is an administered price and the equation just reported describes the behavior of those administering it, the LME price is basically a free price. Hence the equation describing LME price determination is describing the behavior of a market. In effect, one can think of the LME price as adjusting until holders of stocks are satisfied to hold them.[46] In this view, it should be the size of stocks (relative to consumption) rather than the change in stocks that counts for the LME price; however, as one might expect the relationship to be long run, it is sensible to use a distributed lag and also to investigate the effect of lagged stocks as well as current ones.

When this is done, however, a striking result emerges. In every version of the equation tried,[47] what appears to matter is the change in the stocks-consumption ratio and not its level. Indeed, when the change and the level of the ratio are both used in the equation (which is equivalent to using both the current and the lagged values of the ratio), not only is the coefficient of the level very close to zero (and small relative to its asymptotic standard error) but also it is slightly positive, which does not make economic sense.

---

[44] 1967 and 1968 were omitted because of the copper strike. Figures on the change in government stocks of copper required to calculate the change in private stocks are not available before 1949

[45] Elasticity calculations do not make much sense for this equation, since both of the operative variables can (and do) change sign over the period.

[46] There is a vast literature on this subject. See, for example, Telser [14].

[47] There were a number of these They included versions with and without lagged price and with current and lagged stocks-consumption ratios being for year t and $t - 1$ or for year $t - 1$ and $t - 2$. In addition, since stocks figures do not exist for

Denoting the level of Rest-of-World stocks at the end of year t by RWST_t, the two best equations are as follows:[48]

$$\mathrm{USP}_{\mathrm{LME}t} = 285.6 - \underset{\substack{(1565)\\(-2.312)}}{3620}\left(\frac{\mathrm{RWST}_t}{\mathrm{RWC}_t} - \frac{\mathrm{RWST}_{t-1}}{\mathrm{RWC}_{t-1}}\right) + \underset{\substack{(0.4886)\\(1.628)}}{0.7957}\,\mathrm{USP}_{\mathrm{LME}t-1} \quad (32)$$

$$\rho = -0.3 \qquad \text{Years: 1952–1968}$$

and

$$\mathrm{USP}_{\mathrm{LME}t} = 415.2 - \underset{\substack{(738.8)\\(-2.626)}}{1939.7}\left(\frac{\mathrm{RWST}_{t-1}}{\mathrm{RWC}_{t-1}} - \frac{\mathrm{RWST}_{t-2}}{\mathrm{RWC}_{t-2}}\right) + \underset{\substack{(0.1769)\\(3.369)}}{0.5960}\,\mathrm{USP}_{\mathrm{LME}t-1} \quad (33)$$

$$\rho = -0.1 \qquad \text{Years: 1953–1968.}$$

Now, the difficulty in accepting equations such as these is their long-run implications under certain circumstances. In them, the LME price goes down when stocks go up relative to consumption, a perfectly reasonable result. The possibly surprising thing, however, is that if one considers what would eventually happen to price if the stocks-consumption ratio were to remain unchanged indefinitely, the implication is that the LME price would approach an equilibrium level *independent of the level at which the stocks-consumption ratio remained constant.* What the stocks-consumption ratio had been doing before entering the constant phase would matter for the short and middle run, but in the long run, a constant stocks-consumption ratio would lead to the *same* price, whatever

---

the total non-U. S. region, they had to be constructed by cumulating figures for changes in stocks from the estimates given in the next section. This is a satisfactory procedure, given a benchmark estimate of stock level in a particular year, but no such estimate exists except for the United States, where only the change in stocks enters the model in any case. For the rest of the world, stocks were arbitrarily assumed zero at the end of 1949. The only effect of this is in the equation under discussion, where stock levels are divided by consumption. This suggests adding a correction to the equation in the form of the current and lagged reciprocals of consumption. When this was done, the corrections mattered little, but the phenomenon under discussion in the text persisted. Note that the stocks-consumption ratio is very close to the change in stocks divided by consumption, so that the correction should not matter if only changes are important, this slightly reinforces that finding.

[48] The only instrumental variables used in estimating (33) were those required by Fair's method.

the constant level. This is not a result which is instantaneously acceptable, although its importance for short-run or even middle-run forecasting is extremely limited. We can, however, go a long way toward rationalizing it along the following lines.

The LME price not only adjusts according to the desire to hold copper stocks, but, much more fundamentally, it serves as the long-run equilibrator of supply and demand. If, over a very long period, there is no change in the stocks which people desire to hold, then, asymptotically, stocks will not affect the price, which then, asymptotically, approaches the level at which long-run supply and long-run demand are in balance; so that stocks will not in fact change. We have already seen, however, that the supply of copper is extremely elastic in the long run, so that such an asymptotic price, as far as we can predict, is a constant (in real terms) approximating the almost horizontal level of the long-run supply curve.[49] Indeed, if we recall that the change in stocks is *defined* as the difference between total supply and total demand, it becomes apparent that the price adjustment equations merely state that price adjusts in the direction of excess demand and does not move (in the long run) if supply and demand are in long-run balance. It must be emphasized that such arguments and implications are only asymptotic and that speeds of adjustment are slow enough in the copper industry that the long run is very long and not very relevant. Short-run movements in demand (and supply) and consequent short-run changes in stocks will have much more to do with the price over any reasonable horizon than will whatever asymptotic level would be approached were everything forever in equilibrium.

For what it is worth, however, the asymptotic equilibrium level of the LME price is $1,506 per long ton from (32) and $1,164 per long ton from (33) (in 1969 dollars). This compares to an average price of $1,490 (current dollars) in 1969.[50] It must be strongly emphasized that this statement in no way constitutes a prediction that the LME price will approach this level over the next several years. We shall return to this in Section 11.

9. Closing the model

■ Only three more relationships remain to close the model. These are the two identities accounting for changes in stocks and the equation describing net exports of copper from the rest of the world into the United States.

The first two relationships are easy to describe. Let $\Delta USGS_t$ be the change in U. S. government stocks during year t (thousands of metric tons) and let RWX_t denote net exports from the rest of the world to the United States (thousands of metric tons).[51] Then the change in U. S. private stocks is given by the identity

---

[49] This is consistent with the argument in Herfindahl [7], pp. 230–35, that long-run marginal costs in copper supply are nearly constant

[50] It may be worthwhile noting that these equilibrium prices are somewhat above that price which would in a long-run steady state make the U. S. Producer Price and the LME Price the same, using equation (31). Thus our model has a very long-run implication for equilibrium of a producer price below the LME price. This is, of course, not a material issue in terms of the short- or middle-run forecasting ability of the model. We are indebted to Zvi Griliches for commenting on this point.

[51] Excluding scrap already counted in old scrap collections. See above, p. 581.

$$\Delta USS_t \equiv USMP_t + USOS_t + USNS_t + RWX_t - USC_t - \Delta USGS_t, \quad (34)$$

which is to say that the change in U. S. private stocks is the total supply (mine production, old scrap, new scrap, and imports) less consumption and the amount going into U. S. government stocks. Note that new scrap is included in our consumption figures and is netted out of the change in stocks, as it should be, since what matters is copper put through the refining process once during the year less the disappearance of copper into true consumption and government stocks.

A similar identity gives the change in stocks in Rest-of-World. As already remarked,[52] there are no benchmark figures for the size of stocks outside the United States, so we have cumulated the change in stocks from the end of 1949. Therefore, aside from our inability to account for governmental stocks outside the United States,[53] our Rest-of-World stock figures differ from correct ones by an unknown constant. This matters only to the LME price equation, and, as discussed in the preceding section, it matters very little since only changes appear to be important in that equation.

The identity giving the change in Rest-of-World stocks is

$$\Delta RWS_t \equiv ChMP_t + CanMP_t + ZMP_t + RWMP_t + RWS_t - EURC_t - JC_t - RWC_t - RWX_t, \quad (35)$$

which says that the change in Rest-of-World stocks is given by the sum of mine production outside the U. S. (Chile, Canada, Zambia, and Rest-of-World) plus Rest-of-World scrap supply less consumption of copper outside the U. S. (Europe, Japan, and Rest-of-World) less net exports to the United States.

The estimation of the equation explaining net exports to the United States (RWX) is not a simple matter. One naturally expects a prime factor in that equation to be the difference between the producer price and the LME price, but simple attempts to explain net exports by this difference do not get very far. The reason is not hard to find. When the producer price is well below the LME price, there is often rationing of primary copper in the United States and copper may flow into, rather than out of, that country.

This suggests trying several devices. One of these is to divide the difference between producer price and LME price into two variables, depending on whether that difference is positive or negative. When this was done, the results were often suggestive; but when the other variables about to be discussed were added to the equation, the coefficients on the two price-difference variables became almost equal, lending powerful support to the view that the other variables had adequately accounted for the rationing problem.

Those other variables are as follows. The first is the excess of U. S. consumption over U. S. mine production, an indication of the shortfall which must be filled by secondary copper, stock changes, and imports. The second, denoted XD, is a dummy variable representing the presence or absence of export controls in the United States (recall that RWX is *net* exports). This variable was given the value of unity for 1949–1952, 1955–1956, and 1966–1968. For 1953, it was set at 0.75, for 1954 at 0.25, and for 1957 at 0.5 (fractional values occurring when controls were imposed or taken off in the middle of a year). For all other years, the variable was set equal to zero.

---

[52] See note 47 above.

[53] Figures on governmental stocks do not exist The only non-U. S. government stocks whose existence are known to us were held in small amounts by the United Kingdom during the Korean War.

The resulting equation was as follows (note that the prices are in deflated dollars per long ton):

$$\begin{aligned} RWX_t = {} & -795.5 + \underset{\substack{(0.6310)\\(2.214)}}{1.397}\,(USP_{Prod\,t} - USP_{LME\,t}) \\ & + \underset{\substack{(0.4238)\\(2.204)}}{0.9340}\,(USC_t - USMP_t) + \underset{\substack{(55.22)\\(2.640)}}{145.8}\,XD_t \qquad (36) \end{aligned}$$

$$\rho = -0.1 \qquad \text{Years: 1952–1968.}$$

The results are strikingly good, particularly since, as mentioned, attempts to explain net exports without the two final variables lead to very poor results. Elasticity calculations do not mean much when the variables lie close to zero, so we do not give them. An increase in the U. S. producer price of one cent per pound, with the LME price constant, increases net exports to the U. S. by about 31 thousand metric tons, other things being equal. (By way of comparison, net exports were about 201 thousand metric tons in 1963, so there is substantial price sensitivity.) An increase in the gap between U. S. consumption and mine production is filled, other things being equal, about 93 percent by imports into the United States. This figure seems a little high, but (aside from the fact that a more plausible figure would lie within one asymptotic standard error of the point estimate) it must be remembered that this is *not* the same statement as the assertion that 93 percent of the entire gap is filled by imports. Even if the two prices were identical and export controls not present, the large negative constant term would mean that much less of the gap is filled by imports on the average than at the margin.[54]

10. General conclusions

■ In Section 2, we outlined the general way in which the model works. The remainder of this study will discuss specific forecasts. Between these general and specific discussions are the conclusions to be drawn from the model as estimated and reported in the preceding sections. This section briefly treats some of the more significant conclusions.

---

[54] In the estimation of (36), the following instrumental variables were used in addition to those required by the Fair method: lagged mine production (including the United States) and lagged stocks/use outside the United States. Note that XD was taken as exogenous (indeed it is the only exogenous variable in (36)) It is true, of course, that the imposition of export controls is a function of the situation in the U. S. copper market. We assume that it depends on lagged rather than current variables, however.

The overriding fact about the copper market is the very high elasticity of supply coupled with a low adjustment speed and relatively low short-run elasticities. Thus while long-run equilibrium price—that price which would make long-run supply equal long-run demand for current levels of the various exogenous variables—may be well below current price, there is no marked tendency for price to approach such a long-run equilibrium. Steady growth in the activity variables influencing demand, even at a relatively low rate, leads to prices forever above long-run equilibrium because of the inelasticity of short-run supply. In essence, the demand curve is continually shifting outward and new supply coming slowly into production.[55]

On the other hand, this very fact means that prices are sensitive to a reduction in the rate of growth of the exogenous activity variables, especially to a decrease therein. If those variables slacken their growth rate or stop growing altogether, the new supply induced by previously high prices will come on-stream and prices will drop. Since demand is relatively price inelastic, such a drop can be quite substantial. Indeed, it is clear that whereas the high long-run elasticity of supply makes for a stable, if largely irrelevant, long-run price, the low short-run elasticities of supply and demand make for a relatively unstable short-run price which is very sensitive to changes in general economic conditions. This is more true of the LME price than it is of the U. S. producer price, whose fluctuations are deliberately reduced by the price-setters.

Such instability enters the model, of course, through stock accumulation. A fall in demand, given the low short-run and high long-run elasticity of supply, leads to a sharp increase in copper stocks. This in turn depresses the price, which will remain depressed as long as stock accumulation continues. An upturn in general economic conditions, on the other hand, will lead to an outward shift in demand and a decumulation of stocks, providing an upward push on price. Because of the long time required to make adjustments, however, the actual timepath of prices is not a simple one. The remainder of this study discusses forecasts of that path (and of the timepath of the other variables as well) under various assumptions as to Chilean supply.

11. Using the model: forecasts and predictions

■ In this section, we discuss the behavior of the model when it is used to generate predictions of the endogenous variables under various specific numerical assumptions about the exogenous variables. In order to do so, we must both describe those assumptions and discuss the form in which the model was simulated.[56]

Clearly, the forecasts of any econometric model depend heavily on the forecasts of the exogenous variables used therein. Since the latter forecasts are not certain, the forecasts given here for the copper market can only be taken as indicative, even apart from the question of the validity and adequacy of the present model. Further, the reliability of all forecasts decreases as projections move farther into the future. On the other hand, the *differences* in forecasts which result when specific assumptions are changed can be very revealing about the true state of affairs regardless of whether the future values of exogenous variables used throughout are inaccurate and whether the forecasts themselves are in error. This will be important when we discuss the elasticity of demand facing Chile and the effect of new sources of copper supply.

---

[55] This basic result was also found in a preliminary study done some years ago by one of the authors. See Fisher [6]

[56] The computations described in this section were carried out by Nancy Greene using the TROLL system of the National Bureau of Economic Research Computer Research Center.

It is obviously important, however, to use *sensible* forecasts of exogenous variables. In our case, with the period used for estimation ending in 1968, we generated forecasts for 1969–1975. We were able to use actual values of almost all of the exogenous variables for 1969 and 1970. Reasonable forecasts were typically available for 1971 and 1972. For 1973–1975, we made essentially arbitrary assumptions, usually as to constant percentage changes. The details will be found in Appendix 2. Before proceeding, we had to decide exactly what version of the model to use. This involved two minor and two major decisions.

The first decision concerns the fact that there are two alternate versions of the United States demand equation, (24) and (25), and two of the Japanese demand equation, (27) and (28). In both cases, the two equations are distinguished by whether they use an industrial production index or a construction materials index. Collinearity did not permit using both. In the case of Japan, the equation using an index of industrial production seems slightly superior to the one using an index of construction materials: in both cases, it is easier to construct forecasts of the index of industrial production than of the index of construction materials; and letting industrial production be the primary force seems more satisfactory than assigning that role to construction materials (although collinearity suggests that this does not matter very much). Accordingly, in both cases we used the demand equation involving the index of industrial production.

The second problem concerns the Zambian supply equation (13), which (for obvious reasons) is not terribly satisfactory. Since that equation is not very different from a time trend, we replaced it for simulation purposes with such a trend, regressing Zambian mine production on time to obtain

$$\begin{array}{rl} \mathrm{ZMP}_t = 334.4 + & 28.08t \\ & (2.497) \\ & (11.25) \end{array} \qquad (37)$$

$R^2 = 0.9405$ Years: 1954–1957, 1960–1965.

However, perhaps due to political disturbances, Zambian mine production was well below trend in 1969, the first year of our simulation period and the last for which full data were available. Therefore, for 1969, we used actual Zambian mine production, returning to the trend for 1970–1975.

All other endogenous variables were forecast after 1968, actual values for 1968 being used for the lagged terms. A full set of values for the endogenous variables was not available for 1969.

The first major decision concerns the two alternative equations for the LME price, (32) and (33). These equations differ as regards the timing of the stock change affecting the price: in (32) the change is over the same year as the price observation; in (33) it is over the preceding year. There are mild grounds for believing that the former is appropriate—if the latter were, the LME price would be easy to predict and the resulting speculation would destroy the predictive value of the equation—but the argument is not a strong one. Accordingly, we began by experimenting with both versions, which we shall call the "simultaneous" and "lagged" models in order to see which version gave more sensible results.

As it turned out, those experiments were decisively in favor of the simultaneous version. While neither version did a very good job in predicting the LME price over the 1969–1971 period (a matter to which we shall return), the lagged version generated erratic behavior for the U. S. producer price and for imports into the United States. The former variable, in particular, fluctuated while the LME price rose steadily. The reason for this is easy to find. Imports in the model depend on the gap between U. S. demand and U. S. primary supply and on the difference between the two prices [see equation (36)]. In the lagged version, the flow of imports into the United States affects no current price; so that it does not really serve as a method of arbitrage between the two markets. Exports from the United States do not depress and imports into the United States do not increase the current LME price; any such effect follows in the succeeding year as do effects on the U. S. producer price. This is contrary to the relatively free-market nature of the LME and leads to the fluctuating predictions mentioned above. Those fluctuations disappear when the simultaneous version of the model is used and the LME price allowed to adjust currently to diminish the inter-market flow. Accordingly, in what follows, we use the simultaneous version of the model throughout.

Similar experimentation was employed to decide the second major question of whether the estimated autocorrelation structure of the disturbances should be used in simulation and forecasting. The argument in favor of retaining that structure is that it contains important and useful information; not to use it is to ignore important historical regularities. On the other hand, it might be argued that those regularities are merely historical; disturbances are the part of the model we know essentially nothing about, and reliance on the continuance of the same autocorrelation structure is reliance on continued systematic behavior of the unknown. As it turns out, the forecasts are not very different using the two methods and the complete set of results given below are those retaining the estimated autocorrelation structure.

We now turn to the results of the simulation experiments. The first such experiment, the "base" experiment, is simply a set of forecasts using the model as given and assuming that all producers continue to be on their estimated supply curves. The results are given in Table 5.

The first thing to notice about these forecasts is that they are undoubtedly better as one-year projections than as forecasts over several years. This is particularly evident as regards the forecasts of the LME price [frame (d)]. The forecasted LME price (in pounds sterling per long ton) is 619.8 for 1969; the actual average 1969 price was 625. As this represents a rise of 100 over 1968, the fore-

TABLE 5

BASE FORECAST

(a) PRIMARY SUPPLY (THOUSANDS OF METRIC TONS)

YEAR	UNITED STATES	CHILE	ZAMBIA	CANADA	REST–OF–WORLD
1969	1137.7	679.5	719.4	526.1	1516.2
1970	1281.7	700.0	783.7	577.1	1622.3
1971	1376.6	712.4	811.8	603.8	1708.1
1972	1463.8	724.0	839.8	644.3	1789.7
1973	1515.4	736.3	867.9	680.7	1871.7
1974	1568.8	749.2	896.0	722.5	1954.3
1975	1616.5	764.4	924.1	766.4	2042.3

(b) SCRAP (THOUSANDS OF METRIC TONS)

YEAR	UNITED STATES NEW SCRAP	UNITED STATES OLD SCRAP	REST–OF–WORLD SCRAP
1969	909.0	486.4	1949.5
1970	838.0	494.3	1988.0
1971	860.0	500.5	2104.6
1972	882.9	516.3	2216.2
1973	934.2	531.0	2334.6
1974	947.5	546.8	2462.2
1975	995.7	565.5	2606.2

(c) DEMAND (THOUSANDS OF METRIC TONS)

YEAR	UNITED STATES	JAPAN	EUROPE	REST–OF–WORLD
1969	2971.5	1238.6	3167.5	665.7
1970	2806.9	1417.0	3361.3	692.7
1971	2865.1	1568.2	3483.9	713.0
1972	2923.7	1755.6	3617.6	745.9
1973	3053.3	1928.1	3748.4	780.3
1974	3086.9	2115.4	3887.3	815.0
1975	3208.5	2321.6	4033.5	850.1

(d) PRICES

YEAR	LME PRICE (£ / long ton)	UNITED STATES PRODUCER PRICE ¢ / lb	EMJ PRICE ¢ / lb.
1969	619.8	37.3	37.4
1970	678.1	49.1	49.0
1971	638.9	51.3	51.2
1972	679.3	55.4	55.3
1973	736.6	57.2	57.1
1974	798.9	61.7	61.7
1975	877.4	65.7	65.7

(e) EXPORTS (THOUSANDS OF METRIC TONS)

YEAR	REST–OF–WORLD TO UNITED STATES
1969	252
1970	114
1971	9
1972	–12
1973	–24
1974	–72
1975	–78

cast is remarkably good (remember that we began forecasting with 1969). The forecasts for 1970 and 1971 are not good, however. The LME price fell to an average of 597 for 1970 and fell very drastically in 1971, being 528 at the end of the first quarter and 429 at midyear. Our forecast, however, shows a rise in 1970 (to 678.1) and, while we do forecast a fall in 1971, it is nowhere near the correct magnitude. On the other hand, it may be remarked that predictions of the U. S. producer price are quite satisfactory for the same period (but are low for 1969).

What accounts for this? There are several possibilities. In the first place, one naturally expects the model to do better in short-run forecasting. Since the model is dynamic, predictions several years into the future use predictions one or two years into the future as inputs. Thus errors tend to lead to further errors and the model tends to wander. Put more simply, in forecasting 1969, we were able to use actual values of the endogenous variables for 1968 for the lagged endogenous variables in the equations; in forecasting 1970 and later years, on the other hand, we could only so use the forecasted values of those variables for 1969 and later years.

Second, for 1969 (and in nearly all cases for 1970), actual values of the exogenous variables were used in the forecasts. For 1971 and thereafter, only values which were themselves forecasts could be used. As events catch up with such forecasts of the exogenous values, their effect on our own forecasts for the copper market becomes increasingly severe. This was obviously very important in 1971, but even for 1970 we had to use a forecast of the appropriate aluminum price, which may have had a substantial effect.

Third, large new discoveries and the expectation of further ones may have depressed the price in a way not captured by the model. Our supply equations do include new discoveries, but only in a general and systematic way. Sudden success, particularly in new areas, will not be captured in our short-run forecasts. (We examine the effects of the appearance of a large new supplier below.)

Fourth, it is possible that the model correctly predicts underlying tendencies but misses short-run deviations. If one believes this, then one must also believe that the fall in the LME price in 1970–1971 was in some sense an aberration, which is far from clear. Certainly, the model suggests that the general trend is up (recall that the forecasts are in current pounds sterling).

Finally, the model may simply be in error, although its success in fitting the sample period and in predicting 1969 makes one hopeful. In any case, the forecasts of the LME price for 1970–1971 are not good and the model may be at fault. Clearly our equation (30) for predicting the EMJ price of copper is unfortunate and should be modified, since it yields the quite unreasonable implication that this "average" of the LME price and the producer price falls slightly below both. It is not possible, however, to say if this error plays any important role in the results, although this seems unlikely. Only time and further experimentation will tell.

If the model has any validity, even if individual forecasts are in error, comparisons between forecasts using different assumptions may be very revealing. This is easiest to see where the source of the error is in the values of the exogenous variables used; but it is true in other cases as well. Differences between forecasts may be more accurate as reflections of differences between supposedly corresponding actual situations than forecast levels are as reflections of actual levels.

Before exploring such differences, however, a few points are in order regarding the base forecasts of Table 5, taking them at

TABLE 6

ALTERNATE FORECAST 1 CHILEAN SUPPLY UP 10 PERCENT IN EVERY YEAR

(a) PRIMARY SUPPLY (THOUSANDS OF METRIC TONS)

YEAR	UNITED STATES	CHILE AS FORECAST BY MODEL	EXTRA CHILE	ZAMBIA	CANADA	REST-OF-WORLD
1969	1137.5	677.9	67.8	719.4	526.1	1511.9
1970	1268.8	696.5	69.7	783.7	575.3	1612.7
1971	1359.0	708.4	70.8	811.8	600.9	1695.6
1972	1442.5	720.0	72.0	839.8	640.2	1776.0
1973	1495.0	732.7	73.3	867.9	675.9	1857.6
1974	1548.9	746.0	74.6	896.0	717.0	1940.1
1975	1597.7	761.4	76.1	924.1	760.3	2028.1

(b) SCRAP (THOUSANDS OF METRIC TONS)

YEAR	UNITED STATES NEW SCRAP	UNITED STATES OLD SCRAP	REST-OF-WORLD SCRAP
1969	909.0	483.4	1942.2
1970	838.1	491.5	1986.2
1971	864.4	498.8	2104.6
1972	889.3	515.2	2217.5
1973	942.1	530.3	2336.4
1974	955.5	546.1	2463.1
1975	1003.8	565.0	2607.4

(c) DEMAND (THOUSANDS OF METRIC TONS)

YEAR	UNITED STATES	JAPAN	EUROPE	REST-OF-WORLD
1969	2971.5	1238.6	3167.5	665.7
1970	2807.1	1418.6	3366.5	695.0
1971	2876.3	1570.4	3493.6	718.0
1972	2939.7	1757.0	3627.5	751.9
1973	3073.3	1929.0	3757.0	786.4
1974	3107.0	2116.0	3894.1	820.7
1975	3228.9	2322.1	4039.1	855.4

(d) PRICES

YEAR	LME PRICE (£/ long ton)	UNITED STATES PRODUCER PRICE ¢/lb	EMJ PRICE ¢/lb.
1969	610.7	37.3	37.4
1970	665.2	48.1	48.0
1971	630.5	50.6	50.5
1972	673.1	54.6	54.5
1973	732.3	56.7	56.7
1974	794.8	61.2	61.2
1975	873.2	65.3	65.3

(e) EXPORTS (THOUSANDS OF METRIC TONS)

YEAR	REST-OF-WORLD TO UNITED STATES
1969	279
1970	135
1971	42
1972	22
1973	14
1974	–35
1975	–41

face value. The first of these has already been mentioned. According to those forecasts, the outlook over the next few years is for copper shortage with rising prices, particularly for 1972. Second, the gap between the LME price and the U. S. producer price is predicted to narrow (due to the removal of export controls) but not to disappear. Third, and most interesting of all, despite the rising prices, Chilean output is projected as growing at a very slow rate compared with that of any other producer. It seems clear that if the incorrectly high LME price for 1970–1971 were replaced by the much lower actual price, the forecasted behavior of Chilean output would be a decline. *Note that the model forecasts this behavior as part of normal Chilean supply response, independent of unusual political or labor difficulty.* This has clear implications as to the reasons for the actual sluggishness of Chilean output over the last two years and the pressure on the copper companies by the Chilean government.

The next simulation experiment was designed to see the effects of such sluggish Chilean production on prices and on Chilean revenue. In this experiment, we altered the model by assuming that Chilean mine production would be greater than that forecast by the model by 10 percent in every year of 1969–1975. Since the model does not forecast great changes in Chilean output, this is roughly the same as assuming an increase in each year's output of 10 percent of 1969 production. The same set of experiments also serves to show the effect of an increase of the same magnitude by another producing country. The results are presented in Table 6.

The principal feature of these results when compared with the base forecast is seen at once: the assumed increase in Chilean supply does not have much effect on prices. In particular, the behavior of the LME price is not much different from the case of the base forecast. The larger differences, as we should expect, come in the earlier years before the market adjustment. The increase in Chilean supply (over the base forecast) is largely absorbed by decreases in other supplying countries and offset (to a lesser degree) by demand increases.

In other words, Chile is a sufficiently small part of the world market for the demand curve facing her to be quite elastic. This is borne out if we compare the forecasts of Chilean copper revenue generated by the two sets of forecasts. See Table 7. For comparison, we also show forecasted Chilean revenue, on the assumption that the extra output in the alternative forecasts is sold at the LME price by a country other than Chile.

TABLE 7

ALTERNATIVE FORECASTS OF CHILEAN REVENUE
(MILLIONS OF 1957–1959 DOLLARS)

YEAR	BASE FORECAST	ALTERNATIVE FORECAST 1 (CHILEAN PRODUCTION)	ALTERNATIVE FORECAST 1 (OTHER PRODUCTION)
1969	880.7	952.2	865 6
1970	955.7	1026 4	933 1
1971	952 9	1028.4	934.9
1972	990 1	1073.7	976.1
1973	1040.3	1132.1	1029 2
1974	1093 4	1191.4	1083.1
1975	1166 8	1272.4	1156.7

A very rough estimate of the elasticity of the demand facing Chile is about 5½ in the short run and about 8½ in the long run when offsetting reactions in other countries have had time to take place. It is clearly to Chile's advantage to increase her output if

TABLE 8

ALTERNATE FORECAST 2: APPEARANCE OF NEW SUPPLIER EQUAL TO CHILE IN 1973–1975

(a) PRIMARY SUPPLY (THOUSANDS OF METRIC TONS)

YEAR	UNITED STATES	CHILE	NEW SUPPLIER	ZAMBIA	CANADA	REST–OF–WORLD
1969	1137.7	679.5	0	719.4	526.1	1516.2
1970	1281.7	700.0	0	783.7	577.1	1622.3
1971	1376.6	712.4	0	811.8	603.8	1708.1
1972	1463.8	724.0	0	839.8	644.3	1789.7
1973	1513.6	720.6	720.6	867.9	680.5	1837.0
1974	1449.0	716.3	716.3	896.0	704.5	1862.6
1975	1451.2	724.7	724.7	924.1	736.7	1920.0

(b) SCRAP (THOUSANDS OF METRIC TONS)

YEAR	UNITED STATES NEW SCRAP	UNITED STATES OLD SCRAP	REST–OF–WORLD SCRAP
1969	909.0	486.4	1949.5
1970	838.0	494.3	1988.0
1971	860.0	500.5	2104.6
1972	882.9	516.3	2216.2
1973	934.2	501.2	2254.8
1974	948.1	516.8	2428.3
1975	1036.9	547.7	2598.7

(c) DEMAND (THOUSANDS OF METRIC TONS)

YEAR	UNITED STATES	JAPAN	EUROPE	REST–OF–WORLD
1969	2971.5	1238.6	3167.5	685.7
1970	2806.9	1417.0	3361.3	682.7
1971	2865.1	1568.2	3483.9	713.0
1972	2923.7	1755.6	3617.6	745.9
1973	3053.2	1928.1	3748.4	780.3
1974	3088.5	2128.8	3934.1	837.2
1975	3312.5	2340.0	4122.8	897.5

(d) PRICES

YEAR	LME PRICE (£ / long ton)	UNITED STATES PRODUCER PRICE ¢ / lb.	EMJ PRICE ¢ / lb.
1969	619.8	37.3	37.4
1970	678.1	49.1	49.0
1971	638.9	51.3	51.2
1972	679.3	55.4	55.3
1973	642.5	57.2	57.0
1974	663.7	50.0	50.0
1975	772.4	57.6	57.6

(e) EXPORTS (THOUSANDS OF METRIC TONS)

YEAR	REST–OF–WORLD TO UNITED STATES
1969	252
1970	114
1971	9
1972	–12
1973	233
1974	131
1975	261

(f) CHILEAN REVENUE (MILLIONS OF 1957–1959 DOLLARS)

YEAR	CHILEAN REVENUE
1969	880.7
1970	955.7
1971	952.9
1972	990.1
1973	888.0
1974	868.5
1975	973.9

she can do so without creating *unusual* offsetting activities in other producing countries.

Our final simulation covers the case of an "unusual" event, the discovery of a large new source of supply. In this experiment, we assumed that a new supplier began producing (or old ones increased production) suddenly in 1973 at a rate equal to that which the model forecasts for Chile. In other words, this experiment supposed that a new supplier with the Chilean supply curve began producing in 1973. Comparing these results to those of the base forecast should give at least a rough idea of the likely effect of a large new source of supply such as Australia. The results are presented in Table 8, which includes forecasts of Chilean revenue under the supposed circumstances.

As opposed to the results of the first alternative forecast, Table 6, it is clear that these figures differ markedly from those of the base forecast. The appearance of the new supplier causes a fall rather than a large rise in the LME price in 1973 and delays any large recovery and rise until 1975. Even in the latter year, prices are more than 10 percent lower than in the base forecasts (100 pounds sterling per long ton).

More remarkable, however, is the fact that prices still rise substantially in 1975 as other suppliers reduce their outputs, making room for the new ones, and demand responds to past lower prices. This may be taken as another indication that the outlook for the mid-1970s is for reasonably high copper prices. It must be remembered, however, that (1) all these forecasts are predicated on reasonably high rates of growth in the major consuming countries; (2) such predictions are shakiest for 1975; and (3) we make no attempt to forecast past that date, by which time the full effects of the hypothesized new supplier may not have been felt.

Appendix 1

Data used in the model

■ Almost all data on copper were obtained from the annual publication *Metal Statistics*.[57] This journal seems the most reliable, consistent source of data covering the world industry; although more detailed data for the United States alone are available from the Bureau of Mines and the Copper Development Association. The data in *Metal Statistics* are revised frequently and so the most recent figures available were used.[58]

☐ **Prices.** The EMJ and LME prices were taken from the average figure in *Metal Statistics*[59] and the producer price from *Metal Statistics: The Purchasing Guide of the Metal Industries*. The figure was the average producer price of electrolytic copper.[60] Wholesale price indexes were taken from the *United Nations Statistical Yearbook*[61] and exchange rates from *International Financial Statistics*.[62]

---

[57] See [12].

[58] There is one exception to this. The figures given for U. S. direct use of scrap were, through an oversight, not always used in revised form in calculating other variables. The difference is small, however, and experimentation showed the effects to be negligible. In order that the reader may check the definition of the variables used in the model we give, below, page number references to *Metal Statistics, Ibid.*, Vol. 56, 1969.

[59] *Ibid.*, p. 284

[60] See [1], 1970 ed , p 131.

☐ **Primary supply.** Mine production data for the United States, Chile, Canada, and Zambia were taken from *Metal Statistics*.[63] The Rest-of-World figure was the Free World Figure[64] less the countries above. Mine production rather than smelter production was used because some secondary copper is introduced at the smelting stage. Since we estimate secondary production separately, using smelter production would involve double counting.

☐ **Secondary supply.** Total scrap supply for the United States was taken as the sum of production of secondary refined copper plus direct use of scrap.[65] Direct use of scrap includes new and old scrap in good condition which does not require re-refining. For the United States there is a figure for old scrap[66] which we used to estimate the old scrap model. The difference between total scrap supply and old scrap was then used in the new scrap model. Sample figures for 1963 are given in Table 9.

TABLE 9

UNITED STATES SCRAP, 1963 (THOUSANDS OF METRIC TONS)

(1) PRODUCTION OF SECONDARY REFINED COPPER	274.0
(2) DIRECT USE OF SCRAP	738.0
(3) TOTAL SCRAP SUPPLY (1) plus (2)	1012.0
(4) OLD SCRAP	382.7
(5) NEW SCRAP (3) minus (4)	629.3

The figure for item (5) does not correspond to the new scrap figure given in *Metal Statistics*,[67] since the latter excludes direct use of new scrap. For the Rest-of-World (see Table 10) the total scrap supply was defined in the same way as direct use plus secondary

TABLE 10

REST–OF–WORLD SCRAP, 1963 (THOUSANDS OF METRIC TONS)

(1) PRODUCTION OF SECONDARY REFINED	388.0
(2) DIRECT USE	940.9
(3) TOTAL SCRAP SUPPLY (1) plus (2)	1328.9

---

[61] See [15]

[62] The International Monetary Fund. Year-end rates were used, creating a minor inaccuracy.

[63] [12], pp. 190, 196, 198, and 188, respectively.

[64] *Ibid.*, p 19

[65] *Ibid.*, p. 190.

[66] *Ibid.*

[67] *Ibid.*

refined.[68] There is no breakdown into old scrap for countries other than the United States.

In computing the figure for K_t (the stock of copper lying around) we used primary production and net imports of refined;[69] and change in stocks was the computed figure explained below. The net imports of fabricated items were from *Metal Statistics* with each item weighted approximately by copper content.[70] The weights were based on McMahon's figures.[71] See Table 11.

TABLE 11

WEIGHTS USED IN CALCULATING NET IMPORTS OF COPPER FABRICATIONS

COPPER [SCRAP, ALLOY SCRAP (COPPER CONTENT) RODS, TUBES, WIRE, PLATE, OTHER]	1.00
BRASS [RODS, TUBES, WIRE]	0.65
MUNTZ METAL	0.60
ZINC [WIRE, RODS, POWDER]	0.70

Net imports for the Rest-of-World are the negative of the U. S. figure.

☐ **Demand.** The total use of copper figure for each country's demand equation was consumption of refined copper plus direct (or actual) use of scrap.[72]

The exogenous variables used in the demand equations were obtained as follows:[73]

(1) Indexes of industrial production for all countries and groups of countries from the United Nations *Statistical Yearbook.*
(2) German price of aluminum from *Metal Statistics.*
(3) United States index of construction materials from *The Bulletin of the Federal Reserve Board.*
(4) United States inventories of durables from *The Economic Report of the President.*
(5) Japanese index of construction materials from *Monthly Statistics of Japan.*

☐ **Stocks of copper.** Stocks of copper were computed as a residual and the best explanation is provided by the sample figures given in Table 12.

The figures are all from *Metal Statistics* except: (1) government stocks, which came from McMahon and Copper Development Association and (2) imports of copper ores which were from the latter, up to 1953.[74]

A similar computation was made for the Rest-of-World except that figures for government stocks (held by the United Kingdom

---

[68] *Ibid.*, p 25.

[69] *Ibid.*, pp. 190 and 192–93, respectively.

[70] *Ibid.*, p. 194.

[71] See [9].

[72] Figures are given in [12], pp. 24 and 25.

[73] See [15]; [12], p. 277 (average); [5]; *The Economic Report of the President*, and *Monthly Statistics of Japan*, respectively.

[74] See [9], p. 252, and [3], respectively.

TABLE 12

SAMPLE FIGURES FOR THE UNITED STATES, 1963
(THOUSANDS OF METRIC TONS)

(1) MINE PRODUCTION	1100.600
(2) NET IMPORTS OF REFINED, UNREFINED, AND ORES	201.958
(3) OLD SCRAP	382.700
(4) NEW SCRAP	629.300
(5) TOTAL SUPPLY (1) plus (2) plus (3) plus (4)	2314 558
(6) TOTAL USE OF COPPER	2320.400
(7) DECREASE IN STOCKS (6) minus (5)	5.842
(8) DECREASE IN GOVERNMENT STOCKS	10.310
(9) INCREASE IN PRIVATE STOCKS (8) minus (7)	4 468

government for a short period) were not available. Net imports of the Rest-of-World are the negative of the U. S. figure.

The figures for changes in stocks do not provide a benchmark. For the United States the benchmark for private stocks was 614.162 thousand metric tons at the end of 1949. This figure was stocks of copper at primary smelting plants plus refined copper held by fabricators reported in McMahon.[75] For the Rest-of-World, the benchmark was zero at the end of 1949.

Appendix 2

Forecasts used in simulation

■ The exogenous variables in our model are essentially composed of three groups of figures: price indexes and exchange rates; indexes of production; and variables specifically related to the copper industry. We now discuss the forecasts used of each of these.

☐ **Price indexes and exchange rates.** Actual exchange rates were available through 1971. For 1972–1975, exchange rates were assumed constant at end-of-1971 values, assuming that these adequately reflected the effects of the international monetary crisis.

The situation as regards the price indexes for the various countries is more complicated. Here, again, actual figures were available for 1969 and 1970. For 1971 for the United States, Canada, Japan, and Europe and for 1972 for all of these except Europe, we constructed price indexes from the implicit deflators forecast by various econometric models.[76] For the remaining years to 1975—

---

[75] See [9], p. 250.

[76] The forecasts were kindly made available to us by L. R. Klein. The models were: (1) United States. Wharton EFA Mark III Quarterly Model (GNP deflator), (2) Canada: TRACE econometric model of University of Toronto (implicit deflator for consumer expenditure); and (3) Europe: German model of Institute fur Gesellschafts und Wirtschaftswissenschaften of the University of Bonn (implicit GNP

TABLE 13

VALUES OF EXOGENOUS SERIES USED FOR FORECASTING AND SIMULATION

(a) PRICE INDEXES

YEAR	UNITED STATES (1957–1959 = 100)	CANADA (1958 = 100)	CHILE (1958 = 100)	JAPAN (1958 = 100)	EUROPE (1958 = 100)
1969	113.0	122.4	10024.5	109.8	112.8
1970	117.1	124.3	13714.5	113.6	120.7
1971	122.4	127.8	18761.4	120.0	128.0
1972	127.2	131.3	25515.5	126.6	134.4
1973	133.6	135.2	34701.1	132.9	141.1
1974	140.3	139.3	47193.5	139.5	148.2
1975	147.3	143.4	64183.1	146.5	155.6

(b) PRODUCTION INDEXES

YEAR	UNITED STATES (1963 = 100)	EUROPE (1958 = 100)	JAPAN (1955 = 100)	REST–OF–WORLD (1963 = 100)	U.S. DURABLE INVENTORIES (BILLIONS OF DOLLARS)
1969	138.8	186.3	702.2	147.6	63.6
1970	134.7	195.6	814.7	155.8	65.9
1971	138.5	203.5	905.0	162.1	67.1
1972	145.0	211.6	1005.6	167.0	70.2
1973	150.8	220.1	1106.2	173.7	73.7
1974	156.8	228.9	1216.8	180.7	77.4
1975	163.1	238.0	1338.5	187.9	81.3

(c) OTHER VARIABLES

YEAR	EXPORT QUOTA DUMMY	CHANGE IN U.S. GOVT. STOCKS (THOUSANDS OF METRIC TONS)	GERMAN ALUMINUM PRICE DM / METRIC TON
1969	1.00	53.6	226.3
1970	0.75	–6.4	216.5
1971	0.00	0.0	190.3
1972	0.00	0.0	197.7
1973	0.00	0.0	207.6
1974	0.00	0.0	218.0
1975	0.00	0.0	228.9

and for the Chilean price index for 1971–1975—we assumed constant inflation rates more or less consistent with past experience: 5 percent for the United States, Japan, and Europe; 3 percent for Canada, and 36 percent for Chile. Note that we were unable to take account of the anti-inflation program in the United States.

☐ **Production indexes.** Here again actual figures were available through 1970. For the United States and Japan, forecasts for 1971 and 1972 were constructed from the GNP forecasts of the models listed above[77] by assuming a ratio of production index to GNP constant at the 1970 level. For the remaining years and countries, constant growth rates were assumed: 10 percent for Japan; and 4 percent for the United States, Europe, and the Rest-of-World.

Figures on inventories of durable goods in the United States, available through 1970, were constructed from the Wharton EFA

---

deflator) and Belgian model of Free University of Brussels (implicit GNP deflator). In all cases we took the ratio of the desired price index to the available one to be constant from 1970. In the case of Europe, Belgium and Germany were combined with their usual relative weights and the rest of Europe assumed to move proportionally.

[77] See note 76.

forecasts for 1971 and 1972 and taken as constant in real terms thereafter. Basically, both the production indexes and the price indexes show steady growth except for the U. S. recession in 1970.

☐ **Other variables.** Four other exogenous variables remain. Figures on the first of these, the change in U. S. government stocks of copper, are available through 1970. We assumed zero change thereafter. Figures for the German aluminum price were not readily available after 1968. We assumed a constant real dollar price of aluminum. Similarly, we assumed that Chilean policies as to prices received for copper would remain constant from 1968; here, again, events have overtaken the assumption. Finally, export quotas were removed in the United States in September, 1970. We set the dummy variable corresponding to such quotas equal to unity in 1969, 0.75 in 1970, and zero thereafter. Since the relaxation of controls was undoubtedly anticipated, 0.75 may be too high for 1970.[78] The precise values used for all of these variables are shown in Table 13.

References

1. The American Metal Market Company. *Metal Statistics The Purchasing Guide of the Metal Industries.* New York: The American Metal Market Co., annual.
2. Cooper, J P. "Asymptotic Covariance Matrix of Procedures for Linear Regression in the Presence of First-Order Serially Correlated Disturbances." *Econometrica,* forthcoming.
3. Copper Development Association. *Copper Supply & Consumption 1948–1967.* New York: Copper Development Association, 1968.
4. Fair, R. C. "The Estimation of Simultaneous Equation Models with Lagged Endogenous Variables and First Order Serially Correlated Errors." *Econometrica,* Vol. 38, No. 3 (May 1970), pp. 507–16.
5. Federal Reserve System, Board of Governors. *Federal Reserve Bulletin.* Washington, D. C.: Federal Reserve Board, monthly.
6. Fisher, F. M. "The Price of Copper and the Level of Chilean Copper Exports in the Next Decade." Unpublished manuscript. 1962.
7. Herfindahl, O. C *Copper Costs and Prices. 1870–1957.* Baltimore, Md.· Johns Hopkins Press, 1959.
8 Koyck, L. M. *Distributed Lags and Investment Analysis* Amsterdam: North-Holland Publishing Co., 1954.
9. McMahon, A. D *Copper: A Materials Survey.* Washington, D. C.· U. S. Department of the Interior, Bureau of Mines, 1964
10 McNicol, D. "An Econometric Model of the Copper Industry." Unpublished Ph.D. dissertation in progress. Massachusetts Institute of Technology
11. Mamalakis, M. and Reynolds, C. W. *Essays on the Chilean Economy.* Homewood, Ill.. Richard D. Irwin, Inc., 1965
12. Metallgesellschaft Aktiengesellschaft. *Metal Statistics.* Frankfurt am Main: Metallgesellschaft Aktiengesellschaft, annual.
13. Nerlove, M. *The Dynamics of Supply · Estimation of Farmers' Response to Price.* Baltimore, Md.. Johns Hopkins Press, 1958.
14. Telser, L. G "Futures Trading and the Storage of Cotton and Wheat " *Journal of Political Economy,* Vol. 66, No. 3 (June 1958), pp. 233–55.
15 United Nations. *Statistical Yearbook.* New York, annual.

---

[78] We are indebted to David McNicol for assistance here

COMMENTS AND REVIEWS

612 **On the use of β in regulatory proceedings**
WILLIAM J. BREEN AND EUGENE M. LERNER

622 **On the use of β in regulatory proceedings: a comment**
STEWART C. MYERS

628 **Albro Martin's Enterprise Denied: a review**
GEORGE W. HILTON

632 **The determination of the allowed rate of return in a formal regulatory hearing**
PAUL L. JOSKOW

645 **Investment behavior in the U.S. electric utility industry, 1949-1968**
U. SANKAR

On the use of β in regulatory proceedings

William J. Breen
Associate Professor of Finance
Graduate School of Management
Northwestern University

and

Eugene M. Lerner
Professor of Finance
Graduate School of Management
Northwestern University

Public utility regulation requires that a firm's rate of return be linked to its risk, and some analysts have urged that β be used as the appropriate measure of risk. This study explores the use of β in regulatory proceedings and finds it wanting. Empirical measures of β are known to depend upon (1) the estimating equation that is used, (2) the choice of market index, and (3) the specific time period that is selected. Empirical estimates of a firm's β are shown to range from a large positive to a large negative value.

These empirical results are not surprising when it is recognized that corporate β's depend upon managerial actions. Moreover, when managerial policies change, the estimated values of β may be over- or understated. An internal check on the reliability of the estimated β is shown to be given by the sign of α.

1. Introduction

■ Two recent developments in financial analysis bear directly on the problem of the riskiness of an enterprise. The first is the development of a general theory of corporate risk which is based upon portfolio and capital market considerations. The second relates this measure of risk to the corporate financial policies that the firm voluntarily pursues or is constrained to follow by such external forces as regulatory decisions.

One natural application of this theory of risk has been in the problem area of rate regulation of public utilities. Notably, the *Hope* decision required regulatory commissions to relate a utility's allowable return to its risk:[1]

> The return to the equity owner should be commensurate with returns on investments and other enterprises having corresponding risks. That return, moreover, should be sufficient to assure confidence in the financial integrity of the enterprise so as to maintain its credit and attract capital.

Myers has indicated that a regulatory commission could implement the *Hope* decision by basing the utility's allowable rate of return upon its cost of capital.[2] Since this cost is a function of the riskiness of the firm, and risk can be measured within the capital asset pricing model framework, the argument is complete.

---

[1] Federal Power Commission et al. v. Hope Natural Gas Company, 320 U.S. 591 (1949) at 603.

[2] See [6]

The purpose of this paper is to examine both the theoretical and empirical cost of capital estimates that are based upon the capital asset pricing model. Our conclusions may be important within the context of utility rate regulation because we find that the estimates can vary depending upon the estimating equation that is used, the choice of the market index, and the specific time period that is selected. Moreover, corporate decisions influence the riskiness of the firm. Since these decisions are made in the light of an allowable rate of return, the regulator himself can influence the empirical measures of risk that are derived from the capital asset pricing model.

2. Cost of capital and the capital asset pricing model

■ Current financial theory holds that the relevant cost of capital for a firm should be represented by the opportunity cost to the equity holders of the firm. The capital asset pricing model gives a representation of this opportunity cost, in terms of both the expected return to the shareholder and the risk associated with this expected return.[3] In this framework the return offered by any security is be represented as a linear combination of the return offered by a market index, such as the Standard and Poor or New York Stock Exchange index, r_m; a term which represents an individual company effect, α; and a random component, u_i. Thus

$$r_i = \alpha_i + \beta_i r_m + u_i. \tag{1}$$

The variance of return (risk) may then be represented by

$$\mathrm{var}(r_i) = \beta_i^2 \, \mathrm{var}(r_m) + \mathrm{var}(u_i). \tag{2}$$

The first term, $\beta_i^2 \, \mathrm{var}(r_m)$, is called the company's systematic or undiversifiable risk; the second term, $\mathrm{var}(u_i)$, is called the specific or diversifiable risk. When security i is combined with many other securities in a portfolio, diversification substantially reduces the individual company contributions to total risk [$\mathrm{var}(u_i)$]. The nondiversifiable contributions, $\beta_i^2 \, \mathrm{var}(r_m)$, however, cannot be eliminated as long as all companies included in the portfolio have $\beta > 0$.

In equilibrium an optimal portfolio has the property that the risk premium of a company (measured as $ER_i - r_o$, i.e., the expected return to a security minus the pure rate of interest) is proportional to the risk premium in the market as a whole. The factor of proportionality is the coefficient β_i from equation (1). Thus

$$ER_i - r_o = \beta_i (ER_m - r_o). \tag{3}$$

Combining (1) and (3), and assuming that $Eu_i = 0$, we see that one of the implications of the equilibrium assumptions is that $\alpha_i = r_o(1 - \beta_i)$. The return to a security in an equilibrium portfolio can, therefore, be thought of as a linear combination of the pure rate of interest, r_o, and the expected market rate of return ER_m. The weights of these two terms are $1 - \beta_i$ and β_i. Thus β is the "risk levering factor" associated with each individual security.

---

[3] The capital asset pricing model, associated with the work of Markowitz [6], Sharpe [7], Tobin [8], Lintner [4], Fama [3], and others is quite well known in the literature of finance and economics. Hence, we will make use of the assumptions and conclusions of this model without further detailed references or proofs.

The capital asset pricing model therefore gives a representation of the opportunity cost of holding any security in an optimal portfolio in terms of the risk-return combination that an investor can expect to receive.

3. β in theory and practice

■ The capital asset pricing model provides a theoretical link between corporate activities and the equilibrium market return arising from owning the firm's shares. Thus, given the pure rate of interest, the expected return and variance of the return to the market, and the corporate β, the equilibrium set of prices in the capital market may be derived. Since β is the only variable which is dependent on the specific corporation under investigation, it follows that the determination of the specific corporate β is critical to the effective use of the capital asset theory in rate regulation as well as in corporate capital budgeting and individual portfolio management.

Two salient features concerning the corporate β must be stressed. The first is that β is a theoretical construct. It is not directly measurable; rather it is similar to such other useful theoretical constructs as permanent income and demand schedules. Most of the empirical estimates of corporate β to date have been prepared by least squares regression lines through past observations of price changes, as suggested by equations (1) or (3).

From the perspective of regulation, the estimation procedure that is followed to find β should be well understood and the empirical findings should be unambiguous. A well understood procedure implies that there is agreement between technically competent people, and that the estimating equation that is used and the data base from which β is estimated are both appropriate. An unambiguous empirical finding is one that has but a single meaning. The reason for these requirements is simply that the empirical estimates of β that are reached should be reproducible by any interested party. Unfortunately, estimates of β that have been prepared to date do not meet these standards.

The second feature concerning corporate β is that very little empirical work has been done to express β as a function of other measurable corporate variables. One exception is a study by the present authors,[4] in which at several specific time intervals, corporations were ranked on the basis of five fundamental company variables: dividend payout, growth in earnings, stability of earnings growth, market size, and the number of shares traded. These rankings were then used to predict the volatility of the company's stock price in a subsequent period. A population of 1,500 securities was examined at the peak and trough of three stock market cycles. Each security was ranked according to its position with respect to each of the five corporate variables. Companies were ranked from high to low on the basis of number of shares traded and growth rate, and from low to high on the basis of stability of growth rate, market size, and the payout ratio. Companies which ranked in the highest (lowest) x percentile on y number of criteria were then labeled

---

[4] See Breen and Lerner [2].

volatile (stable) with respect to an unweighted index of all companies in the sample. It was shown, for example, that with x equal to 0.25, and y equal to 3, about 80 percent of the volatility (and stability) predictions were accurate; i.e., these companies had β values greater than (less than) 1. The implication of this work for regulatory decision-making is immediate: if a regulator can influence the firm's growth in earnings or the stability of its earning pattern or any of the other three critical variables, it will affect the firm's measured β value.

4. Estimating difficulties

■ **Two equations for estimating β.** The theory outlined in the last section suggests that β may be estimated directly from observed price changes[5] using either equation (1) or equation (3).

Equation (1), however, may be viewed as a more general expression. The reason for this is that equation (3) strictly holds only when the market is in equilibrium. If the market is in equilibrium, and the assumptions of the capital market model are realized, both equations will yield the same estimate of β. β, of course, is a sufficient statistic to capture a firm's return and risk characteristics only when the company is in equilibrium (i.e., if equation (3) holds). If equilibrium does not exist in the equity market, a portion of the firm's total risk (called specific risk) cannot be diversified away by holding a large number of securities in a portfolio. Thus, unless equilibrium holds, β will not capture all the risk that is associated with a portfolio.[6]

The choice of proper relation [equation (1) or equation (3)] for deriving the estimate of β depends upon the assumption that the regulatory commission makes with respect to market equilibrium. From this point of view, equation (1), being more general, seems to be preferred. Estimates of company β's prepared by using both equations will be presented in Section 5, where it will be shown that there is a significant difference between the estimates arising from the two equations.

□ **The time period over which to estimate β.** If the value of β remains stationary over time, then the estimation of β through the use of time-series statistical methods would seem to be in order. Considerations of statistical accuracy would then argue for using as many observations as possible. However, it is doubtful that β values remain stable over long time periods. Strategic and operational management decisions, made in response to changing external or environmental factors, affect a firm's risk characteristics and hence its β value.

When a change does occur in fundamental operating conditions, an efficient equity market will respond with a price change which provides a new equilibrium consistent with the new β value. If such a change occurs during the period of time that is believed to be a homogeneous unit, the measured β will reflect neither the old nor the new true β values but some weighted average of the two.

---

[5] The rate of return is defined as $(\Delta P + D)/P$, where P represents the price of the security and D the dividends received over the period under study. In the empirical work that follows, we ignore the dividends that are received from holding either a specific security or the market average as a whole.

[6] Numerous authors have recently commented about the problems associated with and the accuracy of the estimation of β.

Sample estimates of company β's for various time periods will be presented in Section 5. It will be shown there that these estimates change dramatically as the length of time under study varies, and, further, that the β estimates over a long time period are simply the weighted averages of a series of short-term β estimates.

☐ **The market index used to estimate β and the interpretation of the coefficient.** Market indexes are constructed differently and are not equally volatile. The New York Stock Exchange (NYSE) index, for example is a weighted average of all companies traded on the exchange; the Standard and Poor (S&P) Composite index is a more restricted universe; while an index made up of companies whose financial records are reported on the COMPUSTAT tape is a wider universe. It will be shown that empirical estimates of β can change significantly if different market indexes are used. By significant we mean that within the bounds of statistical accuracy, the estimated β changes from stable (less than one) to volatile (greater than one).

A still further complication with the estimating procedure arises because the size of β depends not only upon the choice of the index that is used, but on the overall direction that the index itself is following. For example, suppose that several changes occur within the corporation over a measurement period and that these have the effect of raising the company's risk premium relative to that of the market, i.e., raising β. The market adjustment process by which a security's risk premium rises is that the price of the stock falls relative to the market. The reason for this is that at the new and relatively lower price, the expected return to a holder of the security will be higher.

Consider two cases. First, assume that during the period in which the risk premium is rising, the general market movement of all security prices is upward. The estimated value of β, derived from the observed price changes, will be less than 1, say 0.5. This estimate, however, will substantially understate the true value of β.

Second, assume that during the period in which the risk premium is rising, the general market index of all security prices is declining. The estimated β value will now be greater than 1, say 1.5. This estimate, however, may overstate the true value of β. Similar remarks are appropriate if a company's equilibrium risk premium declines and the return to the index either rises or falls.

An implication of these remarks is that a test for a shift in β should be made before an individual company's estimate is seriously used. One such test may be performed by observing the calculated α values from equation (1). If the market is in equilibrium, and no change has taken place in the fundamental character of the firm, then both equation (1) and equation (3) hold, and $\alpha = R_o(1 - \beta)$.

The measured α should be positive for companies which are less risky than the market, ($\beta < 1$). However, if the measured β is less than one, and the measured α is negative, a downward shift in the price of the company's stock must have occurred during the measurement period which was unrelated to the general market price movement. If the market is efficient, these price shifts arise as a result of a rise in the underlying β value for the company; the measured β probably understates the true β for such a period. Conversely, if the measured β is greater than one, and the measured α is positive, price increases occurred that were not related to the general market movement of all prices. If the market is efficient, a downward shift in the underlying β occurred; the measured β then overstates the true β. Table 1 serves as a helpful summary of these remarks.

The potential for erroneous judgment in basing decisions on the estimated β is much more severe in terms of individual company analysis, such as rate regulation, than in portfolio selection. From equation (2), we can see that the total portfolio risk (variance) will depend on the β's from several securities. As long as the estimates of β are not systematically biased in one direction, it would seem that for a large enough portfolio, individual errors in the β estimates would tend to offset each other. This, of course, is not true when a decision is made on the basis of an individual β estimate.[7]

TABLE 1

RELIABILITY OF THE ESTIMATED β

		β	
		>1	<1
α	<0	β ESTIMATE PLAUSIBLE	β ESTIMATE UNDERSTATED
	>0	β ESTIMATE OVERSTATED	β ESTIMATE PLAUSIBLE

5. An analysis of empirical results

■ In order to study empirically the theoretical comments made in the previous sections, several different estimates of β were computed for three large companies: International Business Machines Corp., American Telephone and Telegraph Co., and General Motors Corp. In Table 2 we report the estimates of β for (1) varying time periods, (2) different index values (representing the "return to market"), and (3) different estimating equations, i.e., both equations (1) and (3). Using monthly price data beginning October 1964 and ending October 1971, we estimated β for continuous six-month periods, twelve-month periods, and so forth, up to one 84-month period. Both the NYSE index (which is a share-weighted index) and an unweighted price index of all COMPUSTAT industrial and large utility companies were used. The β estimates are calculated using ordinary least squares for equations (1) and (3). The one-month price change relative to the beginning price was used as our measure of rate of return. The interest rate on 30-day treasury bills was used as a proxy for the pure rate of interest when equation (3) was used.

The data in Table 2 show that the length of time used in calculating the estimate of β, the period of time over which the calculation is performed, the index used, and the estimating equation all have substantial effects on the estimate of β. The β estimates for IBM, for example, range from substantially greater than one to a negative value. While this range is greater than that found for GM or AT&T, almost any interpretation of the "riskiness" (and hence its cost of capital) of the companies we studied may be supported by the proper choice of estimation parameters.

It is instructive to examine the estimated α coefficients to test the possibility of shifting β's using the NYSE index as the indicator of the return to the market.

The six-month β estimates for IBM indicate that in early periods β was greater than one, while during later periods it fell to less than

---

[7] The full effects of errors in estimation on portfolio composition are much more complicated than it would seem from simply examining equations (2) and (3). For a discussion of this problem, and some empirical evidence on the exact nature of estimation errors on portfolio composition, see Breen [1]

TABLE 2

EMPIRICAL ESTIMATES OF β

(a) IBM

NUMBER OF MONTHLY INTERVALS	NYSE					COMPUSTAT				
	STRAIGHT LINEAR REGRESSION (EQUATION (1))			EQUILIBRIUM RESTRICTED REGRESSION (EQUATION (3))		STRAIGHT LINEAR REGRESSION (EQUATION (1))			EQUILIBRIUM RESTRICTED REGRESSION (EQUATION (3))	
	TIME	ALPHA	BETA	TIME	BETA	TIME	ALPHA	BETA	TIME	BETA
6	6	0.0079	2.2001	6	2.3628	6	0.0016	0.8957	6	0.9113
	12	0.0251	1.1911	12	1.4104	12	0.0178	0.6735	12	0.8093
	18	0.0394	0.9093	18	1.1575	18	0.0261	0.4407	18	0.7127
	24	0.0398	1.0746	24	0.9479	24	0.0221	0.5826	24	0.6152
	30	0.0847	0.7515	30	1.0545	30	0.0825	0.2901	30	0.7142
	36	0.1241	0.5575	36	0.9717	36	0.1505	−0.1536	36	0.7029
	42	0.1453	0.4886	42	1.0063	42	0.1904	−0.2995	42	0.7086
	48	0.1293	0.3798	48	0.9845	48	0.2007	−0.4761	48	0.6927
	54	0.1384	0.4603	54	1.0165	54	0.1955	−0.3240	54	0.7008
	60	0.1589	0.4745	60	0.8722	60	0.1851	−0.0959	60	0.5989
	66	0.1418	0.8779	66	0.8847	66	0.1226	0.4191	66	0.6303
	72	0.1295	0.7908	72	0.9041	72	0.1207	0.4069	72	0.5956
	78	0.1508	0.4942	78	0.8784	78	0.1622	0.1521	78	0.5891
	84	0.1209	0.6196	84	0.8801	84	0.1186	0.3589	84	0.6107
12	12	0.0116	1.3350	12	1.4104	12	0.0073	0.7616	12	0.8093
	24	0.0197	1.0293	24	0.9479	24	0.0109	0.6039	24	0.6152
	36	0.0589	0.8092	36	0.9717	36	0.0514	0.4276	36	0.7029
	48	0.0566	0.7431	48	0.9845	48	0.0490	0.4206	48	0.6927
	60	0.0754	0.7073	60	0.8722	60	0.0706	0.3471	60	0.5989
	72	0.0640	0.8601	72	0.9041	72	0.0571	0.5120	72	0.5956
	84	0.0568	0.7668	84	0.8801	84	0.0527	0.5033	84	0.6107
18	18	0.0122	1.1373	18	1.1575	18	0.0041	0.6840	18	0.7127
	36	0.0386	0.8876	36	0.9717	36	0.0310	0.5473	36	0.7029
	54	0.0398	0.8856	54	1.0165	54	0.0324	0.5466	54	0.7008
	72	0.0425	0.8829	72	0.9041	72	0.0374	0.5445	72	0.5956
24	24	0.0098	1.0070	24	0.9479	24	0.0054	0.6143	24	0.6152
	48	0.0267	0.8258	48	0.9845	48	0.0195	0.5949	48	0.6927
	72	0.0318	0.8942	72	0.9041	72	0.0278	0.5604	72	0.5956
30	30	0.0156	1.0436	30	1.0545	30	0.0115	0.6739	30	0.7142
	60	0.0292	0.8358	60	0.8722	60	0.0247	0.5247	60	0.5989
36	36	0.0190	0.9635	36	0.9717	36	0.0142	0.6462	36	0.7029
	72	0.0212	0.9055	72	0.9041	72	0.0184	0.5760	72	0.5956
42	42	0.0184	0.9904	42	1.0063	42	0.0140	0.6568	42	0.7086
	84	0.0155	0.8615	84	0.8801	84	0.0139	0.5881	84	0.6107
48	48	0.0130	0.9613		0.9845	48	0.0088	0.6578	48	0.6927
54	54	0.0127	1.0027	54	1.0165	54	0.0092	0.6700	54	0.7008
60	60	0.0145	0.8769	60	0.8722	60	0.0119	0.5745	60	0.5989
66	66	0.0129	0.9141	66	0.8847	66	0.0101	0.6259	66	0.6303
72	72	0.0106	0.9167	72	0.9041	72	0.0091	0.5913	72	0.5956
78	78	0.0099	0.8742	78	0.8784	78	0.0088	0.5766	78	0.5891
84	84	0.0077	0.8795	84	0.8801	84	0.0069	0.6036	84	0.6107

one. The positive estimated α's during the early periods are consistent with this shift, for a positive α and β greater than one indicates that an upward price movement occurred that was not related to the overall price index. This type of price movement will occur if β falls; i.e., when the measured β's overstate the true β.

An examination of the estimated α and β for AT&T yields a

TABLE 2 CONTINUED

(b) GM

NUMBER OF MONTHLY INTERVALS	NYSE					COMPUSTAT				
	STRAIGHT LINEAR REGRESSION (EQUATION (1))			EQUILIBRIUM RESTRICTED REGRESSION (EQUATION (3))		STRAIGHT LINEAR REGRESSION (EQUATION (1))			EQUILIBRIUM RESTRICTED REGRESSION (EQUATION (3))	
	TIME	ALPHA	BETA	TIME	BETA	TIME	ALPHA	BETA	TIME	BETA
6	6	0.0030	1.6254	6	1.6972	6	−0.0051	0.7981	6	0.7284
	12	0.0002	1.6021	12	1.6157	12	−0.0114	0.9563	12	0.8633
	18	−0.0368	1.8187	18	1.5903	18	−0.1278	2.0086	18	0.6199
	24	−0.0369	1.1265	24	1.2423	24	−0.0588	0.8285	24	0.7426
	30	−0.0503	1.6020	30	1.4264	30	−0.1023	1.3429	30	0.8089
	36	−0.0623	1.6940	36	1.4890	36	−0.1694	1.8186	36	0.8461
	42	−0.0762	1.6423	42	1.3739	42	−0.2063	1.9139	42	0.8136
	48	−0.0861	1.7559	48	1.3587	48	−0.2952	2.5044	48	0.7742
	54	−0.0879	1.5321	54	1.1817	54	−0.2461	2.0021	54	0.7029
	60	−0.0778	1.2678	60	1.0743	60	−0.1580	1.2610	60	0.6621
	66	−0.0487	1.0366	66	1.0334	66	−0.0893	0.8181	66	0.6644
	72	−0.0612	1.0362	72	0.9837	72	−0.0899	0.8124	72	0.6723
	78	−0.0630	1.1063	78	0.9459	78	−0.1056	0.9506	78	0.6641
	84	−0.0748	1.1034	84	0.9448	84	−0.1033	0.9060	84	0.6878
12	12	0.0001	1.6034	12	1.6157	12	−0.0047	0.8997	12	0.8633
	24	−0.0183	1.1685	24	1.2423	24	−0.0290	0.7719	24	0.7426
	36	−0.0296	1.5676	36	1.4890	36	−0.0579	1.1643	36	0.8461
	48	−0.0377	1.5140	48	1.3587	48	−0.0721	1.1850	48	0.7742
	60	−0.0369	1.1539	60	1.0743	60	−0.0603	0.8829	60	0.6621
	72	−0.0302	1.0035	72	0.9837	72	−0.0425	0.7342	72	0.6723
	84	−0.0351	1.0122	84	0.9448	84	−0.0459	0.7802	84	0.6878
18	18	−0.0114	1.6057	18	1.5903	18	−0.0199	0.8179	18	0.6199
	36	−0.0194	1.5282	36	1.4890	36	−0.0349	1.0295	36	0.8461
	54	−0.0253	1.2622	54	1.1817	54	−0.0407	0.9064	54	0.7029
	72	−0.0201	0.9928	72	0.9837	72	−0.0279	0.7099	72	0.6723
24	24	−0.0091	1.1892	24	1.2423	24	−0.0144	0.7442	24	0.7426
	48	−0.0177	1.4144	48	1.3587	48	−0.0287	0.9286	48	0.7742
	72	−0.0150	0.9874	72	0.9837	72	−0.0207	0.6981	72	0.6723
30	30	−0.0093	1.4284	30	1.4264	30	−0.0142	0.8672	30	0.8089
	60	−0.0143	1.0911	60	1.0743	60	−0.0211	0.7313	60	0.6621
36	36	−0.0095	1.4901	36	1.4890	36	−0.0159	0.9182	36	0.8461
	72	−0.0100	0.9821	72	0.9837	72	−0.0137	0.6865	72	0.6723
42	42	−0.0096	1.3790	42	1.3739	42	−0.0152	0.8777	42	0.8136
	84	−0.0096	0.9536	84	0.9448	84	−0.0121	0.7063	84	0.6878
48	48	−0.0086	1.3688	48	1.3587	48	−0.0130	0.8359	48	0.7742
54	54	−0.0081	1.1879	54	1.1817	54	−0.0116	0.7511	54	0.7029
60	60	−0.0071	1.0709	60	1.0743	60	−0.0101	0.6888	60	0.6621
66	66	−0.0044	1.0242	66	1.0334	66	−0.0074	0.6675	66	0.6644
72	72	−0.0050	0.9768	72	0.9837	72	−0.0068	0.6751	72	0.6723
78	78	−0.0041	0.9476	78	0.9459	78	−0.0057	0.6744	78	0.6641
84	84	−0.0048	0.9425	84	0.9448	84	−0.0060	0.6928	84	0.6878

different picture. The estimated β coefficients for AT&T are consistently less than one. The estimated α's, contrary to the implication of equation (4), are typically negative. Within the framework of the equilibrium capital market model, this development, which reflects a sharp decline in the prices of the stock relative to the market, could occur if the risk premium for AT&T was rising, i.e., β was

TABLE 2 CONTINUED

(c) A T & T

NUMBER OF MONTHLY INTERVALS	NYSE					COMPUSTAT				
	STRAIGHT LINEAR REGRESSION (EQUATION (1))			EQUILIBRIUM RESTRICTED REGRESSION (EQUATION (3))		STRAIGHT LINEAR REGRESSION (EQUATION (1))			EQUILIBRIUM RESTRICTED REGRESSION (EQUATION (3))	
	TIME	ALPHA	BETA	TIME	BETA	TIME	ALPHA	BETA	TIME	BETA
6	6	−0 0043	0 5970	6	0.5197	6	−0.0058	0 2357	6	0.1358
	12	−0.0107	0.4168	12	0 3132	12	−0 0147	0.2737	12	0 1385
	18	−0.0392	0 7474	18	0.4953	18	−0 0926	1 1059	18	0.0781
	24	−0 0240	0.6325	24	0.7118	24	−0.0326	0.2283	24	0.1804
	30	−0.0375	0.7754	30	0.6362	30	−0 0520	0.4870	30	0 2039
	36	−0.0650	0 8043	36	0 5811	36	−0 1159	0 8631	36	0 1853
	42	−0 0711	0 7174	42	0.4583	42	−0.1331	0 8895	42	0.1680
	48	−0.0654	0.7606	48	0.4474	48	−0.1521	1 0508	48	0.1455
	54	−0 0595	0 7769	54	0.5314	54	−0 1190	0.8321	54	0.1926
	60	−0 0638	0 7000	60	0.5381	60	−0 0999	0 6130	60	0.2288
	66	−0.0528	0.5850	66	0.5854	66	−0.0738	0 4276	66	0 3009
	72	−0 0782	0.6741	72	0.6090	72	−0.0945	0.4900	72	0.3433
	78	−0.0891	0.8125	78	0 5849	78	−0.1193	0.6886	78	0.3637
	84	−0.1063	0.8201	84	0 5923	84	−0.1240	0.6355	84	0.3719
12	12	−0 0049	0 3552	12	0 3132	12	−0 0060	0 2010	12	0.1385
	24	−0.0119	0.6599	24	0.7118	24	−0 0161	0.1969	24	0 1804
	36	−0 0309	0 6723	36	0.5811	36	−0 0396	0 4157	36	0.1853
	48	−0.0288	0 5770	48	0.4474	48	−0 0371	0 3711	48	0.1455
	60	−0 0303	0.6065	60	0.5381	60	−0 0381	0 3740	60	0.2288
	72	−0.0387	0.6823	72	0 6090	72	−0 0447	0 4077	72	0 3433
	84	−0.0499	0.6906	84	0.5923	84	−0 0551	0 4845	84	0.3719
18	18	−0 0121	0.5205	18	0 4953	18	−0 0145	0 2429	18	0.0781
	36	−0.0202	0.6312	36	0.5811	36	−0 0239	0.3235	36	0.1853
	54	−0 0171	0.5941	54	0.5314	54	−0 0197	0.3021	54	0.1926
	72	−0 0257	0.6186	72	0.6090	72	−0 0293	0 3822	72	0 3433
24	24	−0.0059	0 6733	24	0 7118	24	−0 0080	0.1815	24	0 1804
	48	−0.0135	0.5012	48	0.4474	48	−0.0148	0 2390	48	0.1455
	72	−0.0192	0 6118	72	0.6090	72	−0 0218	0.3698	72	0 3433
30	30	−0 0069	0.6460	30	0.6362	30	−0 0072	0.2450	30	0 2039
	60	−0.0117	0 5550	60	0.5381	60	−0 0133	0 2782	60	0.2288
36	36	−0.0100	0.5915	36	0.5811	36	−0 0109	0 2474	36	0 1853
	72	−0.0128	0 6050	72	0 6090	72	−0.0144	0.3576	72	0 3433
42	42	−0.0090	0 4720	42	0.4583	42	−0.0098	0 2212	42	0 1680
	84	−0.0137	0.6073	84	0.5923	84	−0.0146	0.3958	84	0 3719
48	48	−0 0065	0.4666	48	0.4474	48	−0.0067	0.1913	48	0.1455
54	54	−0 0055	0 5438	54	0.5314	54	−0.0056	0.2270	54	0.1926
60	60	−0.0058	0.5384	60	0.5381	60	−0 0064	0.2513	60	0 2288
66	66	−0.0048	0.5716	66	0.5854	66	−0.0061	0.3032	66	0.3009
72	72	−0 0064	0.5982	72	0 6090	72	−0.0071	0.3433	72	0 3456
78	78	−0.0059	0.5879	78	0.5849	78	−0.0065	0.3763	78	0 3637
84	84	−0.0068	0.5914	84	0.5923	84	−0.0072	0.3796	84	0.3719

increasing. The measured value of β, therefore, consistently understates the true value of β for AT&T throughout the entire period.

The six-month β estimates for GM are all greater than one and the α estimates are almost always negative. These results are consistent with the view that the β estimates are approximately accurate.

6. Implications for rate setting

■ It is proper to link allowable rates of return to the firm's cost of capital and the firm's cost of capital to risk. Both the theory and the empirical work presented above, however, indicate that a regulatory body can be misled if it attempts to use estimated β coefficients from time series regression analysis in an uncritical way for rate regulation.

If the capital asset pricing model is to be used as the underlying framework for valuation, the relation between the corporate activities, both new and old, and the relative risk of the corporation must be examined in much more detail than has been done to date. It is reasonable to believe that such examinations will reveal that regulatory decisions themselves directly effect the value of β, for they influence the corporation's growth rate, stability, size, and payout. If this conjecture is borne out, the regulatory body must be prepared to assess not only the impact of its decisions on the allowable return, but also its effect on the riskiness of the company.

References

1. Breen, W. J "The Effects of Estimation Errors on Portfolio Composition." Unpublished manuscript. 1972
2. ——— and Lerner, E M. "Corporate Financial Strategies and Market Measures of Risk and Return " Unpublished manuscript. 1972
3. Fama, E F "Risk, Return and Equilibrium: Some Clarifying Comments." *Journal of Finance,* Vol 23, No. 1 (March 1968), pp. 29–40
4. Lintner, J "The Valuation of Risky Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *The Review of Economics and Statistics,* Vol 47, No. 1 (February 1965), pp. 13–37
5. Markowitz, H *Portfolio Selection: Efficient Diversification of Investments* New York: John Wiley and Sons, Inc., 1959
6 Myers, S. C. "The Application of Finance Theory to Public Utility Rate Cases." *The Bell Journal of Economics and Management Science,* Vol 3, No. 1 (Spring 1972), pp. 58–97.
7. Sharpe, W. F. *Portfolio Theory and Capital Markets.* New York: McGraw-Hill Book Co., 1970.
8 Tobin, J. "Liquidity Preference as Behavior toward Risk." *Review of Economic Studies,* Vol. 25, No. 1 (February 1958), pp. 65–86.

On the use of β in regulatory proceedings: a comment

Stewart C. Myers
Associate Professor of Finance
Sloan School of Management
Massachusetts Institute of Technology

The purpose of this comment is to provide further clarification of the usefulness of β as a measure of risk in the regulatory process. Breen and Lerner list several of the problems in using this coefficient and argue that a great deal of caution is necessary when interpreting its meaning, but they fail to provide sufficient or correct reasons for their concern. Although the comment is critical of the Breen-Lerner paper, it also redirects attention to the real problems associated with using the β coefficient in regulation.

1. Introduction

■ There *are* problems in using β's as risk measures in regulatory proceedings. Caution is certainly called for. However, Breen and Lerner are cautious for several wrong reasons. They remind me of the proverbial chicken, who worried not about the fox nor the farmer's stewpot, but about whether the sky was falling.

The purpose of this comment is to remove certain items from Breen and Lerner's list of problems, and thereby to give a clearer picture of the usefulness of β in the regulatory arena. However, the case in favor of using β will not be treated in any detail.

I will use Breen and Lerner's notation and equation and reference numbers throughout.

2. The capital asset pricing model *ex ante* and *ex post*

■ Suppose the market for stocks is always in equilibrium over a period of time, and that the assumptions underlying the capital asset pricing model hold. This means that equation (3) holds at the beginning of each subperiod *t in terms of expectations:*

$$Er_{it} - r_{ot} = \beta_{it}(Er_{mt} - r_{ot}). \tag{3}$$

Alternatively, we have

$$Er_{it} = \alpha_{it} + \beta_{it} Er_{mt}, \tag{3a}$$

where $\alpha_{it} \equiv r_{ot}(1 - \beta_{it})$ and $\beta_{it} = \mathrm{Cov}(r_{it}, r_{mt})/\mathrm{Var}(r_{mt})$.

The capital asset pricing model does *not* imply that expectations are necessarily realized *ex post* during the subperiods. Assuming that equations (3) and (3a) hold at the start of t, then *ex post* we must introduce an "error" term, V_{it}:

$$r_{it} = \alpha_{it} + \beta_{it} r_{mt} + V_{it}. \tag{4}$$

Comparing equations (3) and (4) it is obvious that $EV_{it} = 0$ if $r_{mt} = Er_{mt}$. However, if r_{mt} differs *ex post* from Er_{mt}, then Er_{it} may not equal the "predicted" value. For example, if the distribution of the error term V_{it} is skewed, and it is correlated with r_{mt}, then $Er_{it} | r_{mt}$ will not be equal to $\alpha_{it} + \beta_{it} r_{mt}$.

The capital asset pricing model cannot be tested *ex post* without specifying the stochastic process generating *ex post* returns. In this paper I will assume that r_{it} and r_{mt} are drawn from a bivariate normal distribution.[1] This is sufficient for the discussion to follow.

Let us assume further that β_{it} and r_{ot} are constant over time, and that (3) holds at the start of each subperiod. From (4), the average return over a sequence of T subperiods is

$$\bar{r}_i = \alpha_i + \beta_i \bar{r}_m + \bar{V}_i, \tag{5}$$

where $\bar{r}_i = \Sigma r_{it}/T$, $\bar{r}_m = \Sigma r_{mt}/T$, and $\bar{V}_i = \Sigma V_{it}/T$.

The capital asset pricing model says that $EV_{it} = 0$ and thus $E\bar{V}_i = 0$, assuming that r_{it} and r_{mt} are drawn from a bivariate normal distribution. But this does not require that $\bar{V}_i = 0$ for a given stock and a given period of time.

Equation (5) is specified in terms of the true values of β_i and α_i. It is not the equation of a regression line fitted to the *ex post* observations. The regression line is a least squares fit of[2]

$$r_{it} = \hat{\alpha}_i + \hat{\beta}_i r_{im} + u_{it}, \tag{1}$$

where the hats (^) indicate estimated values. This distinction is important. Since the least squares estimate forces $\Sigma u_{it}/T = 0$, and the fitted line passes through the point of means $(\bar{r}_i, \bar{r}_m)$, we have

$$\bar{r}_i = \hat{\alpha}_i + \hat{\beta}_i \bar{r}_m. \tag{6}$$

Therefore, using (5), we obtain

$$\hat{\alpha}_i = \alpha_i + (\beta_i - \hat{\beta}_i)\bar{r}_m + \bar{V}_i.$$

Thus, although $E(\hat{\alpha}_i) = \alpha_i$, the odds are small that $\hat{\alpha}_i$ will be exactly equal to α_i for a given stock and a given period. The value of $\hat{\alpha}_i$ will be affected if expectations are not realized ($\bar{V}_i \neq 0$) or if $\hat{\beta}_i$ turns out to be an over- or under-estimate of β_i.

This rather elementary discussion leads directly to two important points about the Breen-Lerner paper.

First, we should forget about the "equilibrium-restricted regressions." It is not proper to set the constraint, $\hat{\alpha}_i = (1 - \hat{\beta}_i)r_o$. Doing so assumes not only equilibrium *ex ante* but also that expectations are realized *ex post* (i.e., $\bar{V}_i = 0$). Further, the likely effect of doing so is to introduce an error in $\hat{\beta}_i$, since forcing

$$\hat{\alpha}_i - \alpha_i = (\beta_i - \hat{\beta}_i)\bar{r}_m + \bar{V}_i = 0$$

also forces

$$(\hat{\beta}_i - \beta_i)r_m = \bar{V}_i.$$

Breen and Lerner apparently have a concept of "*ex post* equilibrium" in mind, but this is meaningless in terms of the capital asset pricing model. They are right in pointing out[3] that β cannot be a sufficient risk measure if the market is not in equilibrium—but, again, this is equilibrium *ex ante*.

---

[1] See Roll [10], p 272

[2] It is not clear whether Breen and Lerner interpret (1) as a regression equation or as an assumption about the stochastic process generating security returns. (In the latter case, it would be used in the sense of the "diagonal" or "single-index" model of portfolio selection. See Sharpe [7], pp. 118–22.) In any event, (3) is not derived from (1).

In this paper, (1) is interpreted as a regression equation.

Second, we should forget about using fitted values $\hat{\alpha}_i$ as a basis for detecting instability in β_i. Even if (3) always holds, and the true β_i and α_i are constant, $\hat{\alpha}_i$ may still differ from $(1 - \hat{\beta}_i)r_o$ if $\hat{\beta}_i \neq \beta_i$ or $\bar{V}_i \neq 0$. Of course it is also possible that the observation of an $\hat{\alpha}_i$ which differs from $(1 - \hat{\beta}_i)r_o$ reflects an unanticipated shift in β_i during the period under examination. But Breen and Lerner do not show why fluctuations in $\hat{\alpha}_i$ should be attributed solely to this latter reason.

It is conceivable that shifting β's are in fact the major determinant of high or low $\hat{\alpha}$'s, even though this does not follow from the theory. But if the argument is put on an empirical basis there must be empirical support. Breen and Lerner would have to show, with a broad sample of stocks, that the difference between $\hat{\alpha}$ and $r_o(1 - \hat{\beta}_i)$ could be used to predict $\hat{\beta}$'s estimated from subsequent data. But there is no indication in the paper that they are contemplating such tests. They do not even tell us whether the $\hat{\alpha}$'s they observe are significantly different from the predicted values.

3. Choice of the market index

■ Breen and Lerner note that the size of the measured $\hat{\beta}$'s reflects the specific index chosen to represent r_m, and they allege that this is a problem. Is it really?

Conceivably analysts could be fooled by different β's due solely to different indexes. But apart from such naiveté, which will doubtless be eliminated by the Breen-Lerner article, the choice of an index should not lead to overwhelming difficulties.

In the first place, not all indexes are created equal from the point of view of the capital asset pricing model. The theory specifies r_m as the return on a market portfolio composed of all shares.[4] The appropriate index should be constructed from observed returns on all stocks, weighted by the relative market values of the outstanding shares of each stock.

There is no available index which measures r_m exactly. But this is not serious so long as the indexes are highly correlated with each other and with the true market return r_m.

To see this, we can start with (3) and (4), which are stated in terms of the true values of α_i, β_i, and r_m. For simplicity, let us ignore measurement errors in $\hat{\beta}_i$, so that if (1) could be fitted to the true index, then we would find

$$\hat{\beta}_i = \beta_i = \frac{\mathrm{Cov}(r_m, r_i)}{\mathrm{Var}(r_m)}.$$

If another index r_m^* is used, a different β_i^* will be found:

$$\beta_i^* = \frac{\mathrm{Cov}(r_m^*, r_i)}{\mathrm{Var}(r_m^*)}.$$

---

[3] See p. 615 above.

[4] In principle, there is no reason why the index should not be expanded to include all assets—not only stocks but government bonds, real estate, etc. Most work so far has been restricted to the stock market, however.

Let us suppose, however, that r_m^* is perfectly correlated with r_m, although it may be more or less volatile.[5] This means that $r_m^* = \gamma_0 + \gamma r_m$, which, in turn, implies that $\mathrm{Cov}(r_m^*, r_i) = \gamma\,\mathrm{Cov}(r_m, r_i)$, $\mathrm{Var}(r_m^*) = \gamma^2\,\mathrm{Var}(r_m)$, and therefore that

$$\beta_i^* = \frac{\gamma\,\mathrm{Cov}(r_m, r_i)}{\gamma^2\,\mathrm{Var}(r_m)} = \frac{\beta_i}{\gamma}. \tag{7}$$

Equation (7) implies that, for any two securities, $\beta_i^*/\beta_j^* = \beta_i/\beta_j$. Thus, the use of r_m^* does not cause any error in assessing the relative risk of the two securities.

Let us suppose the analyst wants to go one step further, and obtain an estimate of Er_i via the capital asset pricing model. (I am not necessarily recommending this.) Ideally, he would estimate Er_m and β_i, observe r_o, and compute Er_i from (3). Instead he estimates Er_m^* and β_i^* and computes $r_o + \beta_i^*(Er_m^* - r_o)$. But this is in fact the right answer. If $\beta_i^* = \beta_i/\gamma$, then[6] $Er_m^* - r_o = \gamma(Er_m - r_o)$, and

$$\beta_i^*(Er_m^* - r_o) = (\beta_i/\gamma)\gamma(Er_m - r_o) = Er_i. \tag{8}$$

In summary, we have found that the index chosen is not critical to the tasks of judging firms' relative risks or computing their cost of capital. (Of course, these tasks may be difficult for other reasons.) It is true that one might accomplish these tasks and still not know whether the firm in question is "more or less risky than the market as a whole;"[7] but this does not seem to be so important.

This argument does depend on the assumption that the indexes used are good proxies for (i.e., highly correlated with) the market return r_m. This cannot be proved absolutely, but it is hard to reject such broadly based indexes as the Standard and Poor's 500 or the New York Stock Exchange Index as suitable proxies. For one thing, the securities included in such indexes comprise a large fraction of the *value* of the market portfolio. That is, the stocks excluded are those of small firms. Also, the fact that the returns on such indexes are highly intercorrelated[8] is consistent with the view that each is highly correlated with the true market return.

4. Stability of β

■ Breen and Lerner's remaining major point is that β is not stable over time and that regulatory decisions may increase the instability.

The capital asset pricing model does not require that β is stable over time. However, we hope that firms' β's are reasonably stable, for otherwise the model is not operationally useful. Of course the definition of "reasonable stability" depends on the application one

---

[5] The assumption of perfect correlation of r_m and r_m^* is not strictly necessary to the following discussion. In the absence of perfect correlation, we have

$$r_m^* = \gamma_o + \gamma r_m + \epsilon$$

Equations (7) and (8) follow as long as the error term ϵ is independent of r_m, r_i, and r_j. Of course the usefulness of the estimated $\hat{\beta}_i^*$ decreases as $\mathrm{Var}(\epsilon)$ increases

[6] Since (3) holds for each security included in the index, it holds for the index also. Thus $Er_m^* - r_o = \beta r_m^*(Er_m - r_o)$, where $\beta r_m^* = \mathrm{Cov}(r_m^*, r_m)/\mathrm{Var}(r_m)$ If $r_m^* = \gamma_o + \gamma r_m + \epsilon$, and ϵ is independent of r_m, then $\beta r_m^* = \gamma$.

[7] Breen and Lerner, note 5.

[8] See Sharpe [7], p. 148.

has in mind; but there must be very few feasible applications if β's are really as unstable as the figures Breen and Lerner present. They conclude[9] that "almost any interpretation of the 'riskiness'... of the companies ... we studied may be supported by the proper choice of estimation parameters."

This is misleading. Will anyone really be foolhardy enough to find an estimated $\hat{\beta}$, using six months' data, and to assume that this estimate reveals anything about the risk of the company in question? Any such person will be demolished by a simple question: "What is the standard error of your estimated $\hat{\beta}$?"

I do not mean to make light of the problem of stability of β or of the problem of errors in the estimated $\hat{\beta}$. It is just that Breen and Lerner have not shed any new light on these problems. In fact, they may have confused matters by not making a clear distinction between fluctuations in the estimated $\hat{\beta}$ and possible shifts in the true β. They should have asked whether the estimated $\hat{\beta}$'s, considered in light of their standard errors, are sufficient evidence for rejecting the hypothesis that the true β is stable.

Breen and Lerner are correct in pointing out that regulatory decisions can affect utilities' risks. Further work is needed to identify situations in which the effect is empirically significant and to devise ways of taking the effect into account in such situations.

5. Summary

■ Breen and Lerner say that they have explored the use of β in regulatory proceedings and found it wanting.[10] However, these explorations have not been productive and thus do not support this conclusion.

Specifically, I have shown[11] that:

(1) There is no theoretical basis for "equilibrium restricted regressions."
(2) Their proposal to measure shifts in β by observed $\hat{\alpha}$'s has not been shown to be useful.
(3) The choice of the market index is not a major source of difficulty.
(4) They have not distinguished sampling errors in $\hat{\beta}$ from shifts in the true β, and have very likely exaggerated the instability of the latter.

The point of these comments is not just to criticize the Breen-Lerner paper, but to refocus attention on the *real* problems in using β in a regulatory proceeding, which are as follows.

First, β cannot be measured precisely. The possible errors in $\hat{\beta}$ limit the precision of the conclusions that can be drawn.

Second, β may not be stable. This may also limit the precision of any conclusions, unless ways can be found to explain and predict shifts in β.

Third, the capital asset pricing model may not be the whole story about risk and return, on either a theoretical or an empirical basis.[12] It would be foolish, given the present state of knowledge, to propose (3) as a complete basis for regulation. On the other hand, that fact that the model may not be exactly true does not mean that it should be thrown away. A good case can be made for the use of $\hat{\beta}$'s as part of the evidence in regulatory proceedings. But time and space are not sufficient to make that case here.

---

[9] See p. 617 above.

[10] See p. 612 above.

[11] I hasten to add that the results in this comment are basically a restatement of standard theory and not original.

References

1–8. As listed in W J. Breen and E M. Lerner, "On the Use of β in Regulatory Proceedings," pp. 612–21 above.

9 Jensen, M. C "The Foundations and Current State of Capital Market Theory." Forthcoming in M. C. Jensen, ed., *Studies in the Theory of Capital Markets,* New York: Praeger Publishers, 1972.

10. Roll, R. "Bias in Fitting the Sharpe Model to Time Series Data " *Journal of Financial and Quantitative Analysis,* Vol. 4, No. 3 (September 1969), pp. 271–90.

---

[12] See Jensen's recent review of theoretical and empirical work on the capital asset pricing model [9].

Albro Martin's *Enterprise Denied*

Reviewed by
George W. Hilton
Professor of Economics
University of California at Los Angeles

■ This volume,[1] a recent doctoral dissertation in history at Columbia, is an effort to identify the origin of the decline of the railroads in the expansion of the Interstate Commerce Commission's statutory body of authority in the "Progressive" period, and in the Commission's policy toward overall rate increases between 1910 and 1914. Specifically, Martin argues that the Hepburn Act of 1906 and the Mann-Elkins Act of 1910 represented a triumph of "archaic Progressivism;" that the two acts, through establishing maximum rate regulation and investigation-and-suspension procedures in rate actions, made possible a depression of the railroads' rate of return and induced a capital starvation after 1908 that rendered the railroads unable to meet the peak demands of World War I or to deal with the rivalry of the carriers which arose after 1920. Worse, in the author's view, the repressive nature of the regulation inhibited the spirit of enterprise among railroad executives so as to make them slow to respond to prospective changes in technology or other dynamic conditions.

Martin considers this a revisionist interpretation of the history of the period; it has less novelty than he appears to recognize. John W. Barriger, variously president of the Monon, Pittsburgh & Lake Erie and Missouri-Kansas-Texas railroads, has vigorously argued this interpretation of railroad history for some 40 years, but Martin fails to cite his writings.

Although this book won the first Columbia University Prize in American Economic History in Honor of Allan Nevins, it impresses me as a bad example of economic analysis and, less defensibly, a poor piece of historical scholarship. On the whole it amounts to a muddled treatment and a general misinterpretation of the origin of the decline of the railroads.

The book's deficiencies stem directly from the author's methodology. Apparently, Martin gained an acquaintance with Marshallian analysis on an intermediate-price-theory level, but rejected it as inappropriate on the grounds associated with the followers of Thorstein Veblen and the older generation of institutional labor economists: an *a priori* deductive system is inadequately related to reality and likely to lead to erroneous conclusions. Similarly, he rejects the counterfactual method of economic history particularly identified with R. W. Fogel's *Railroads and American Economic Growth*.[2] Martin is left with a traditional methodology of inductive scholarship which he carries on in the fashion of many other historians of the Progressive period: identification of political pressures and their various philosophic bases, relation of the political action to the overall conception of Progressivism, ostentatious show of familiarity with the sources, and quotations from Mr. Dooley as required.

---

[1] A. Martin, *Enterprise Denied: Origins of the Decline of the American Railroads, 1897–1917* (New York: Columbia Univ. Press, 1971).

[2] R. W. Fogel, *Railroads and American Economic Growth: Essays in Econometric History* (Baltimore: Johns Hopkins Press, 1964).

Martin's antitheoretical methodology inevitably prevents him from dealing with the ICC within the context of cartel theory, and thus precludes him from seeing the Commission for what it was and is [p. 44]:

> Under the regulatory system which the nation established in 1906, the ICC was to perform the role of ombudsman. It was not to usurp the actual rate-making function, but to confine itself to relieving individuals from the indifference, incompetence and malice which in the real world frequently impede the smooth workings of economic theory

Gabriel Kolko and others in the 1960s demonstrated that the original purpose of the ICC was stabilization of railroad cartels which had proven themselves intolerably unstable under the legal devices open to private management within a framework of jurisprudence hostile to cartelization. Martin indicates acquaintance with this literature, but in Kolko's case, explicitly rejects the interpretation. As a consequence, Martin is writing about a short period in a long history of cartelization without an adequate conceptual scheme or adequate basis in historical scholarship either in the earlier or later periods. Had he devoted himself in any detail to the Elkins Act of 1903, which is within the period he chose, he could hardly have failed to see the overall framework of regulation for what it was. As Kolko demonstrated, this statute, which imposed stringent penalties on railroads and shippers alike for rebating, was drafted in the legal department of the Pennsylvania Railroad, and could have served only to facilitate enforcement of collusively set rates.

Martin devotes approximately the first third of his book to a demonstration of what he considers to be the vitality of the industry prior to the repressive legislation of 1906 and 1910. His description of improvements in technology is valid enough, but consistent with his failure to use cartel theory, he interprets continued building of railroads in the early twentieth century as evidence of the industry's vitality. The Union Pacific's Salt Lake City-Los Angeles line of 1906 was a defensible expansion of the national network, but the Western Pacific and the Milwaukee Road's Puget Sound extension of 1909 were straightforward examples of the "unnecessary railroads" parallel to existing lines which the cartelization gave an incentive to build.

Martin treats the acts of 1906 and 1910 mainly by reporting on the political discussions which led to passage, rather than by discussing either the consistency of the provisions or their specific consequences on the organization of the industry. Thus he is able to conclude that the inauguration of maximum rate regulation in the Hepburn Act was a repressive measure. Actually, even though the rates set were maximum rates, the fact that they had the authority of statutes while in force made them easier to establish and enforce collusively, and in consequence, the Hepburn Act contributed greatly to the stabilization of the cartels. The same statute's restrictions on pass issuance and "accessorial services" were inhibitions on clandestine rate shading. The investigation and suspension procedures established in 1910 and recognized for decades, were a powerful inhibition to promiscuous rate reduction, and the Mann-Elkins Act's revision in Section 4 of the Act of 1887 restored its effectiveness against the practice of charging more for a shorter haul than for a longer haul. Without an effective Section 4, the Commission was unable to put down rate wars in which a railroad cut rates between points which it served in rivalry to parallel railroads below the level of rates to intermediate points.

Basically, what the legislation of 1903, 1906, and 1910 did was rectify the adverse judicial decisions of the 1890s and otherwise patch up the Commission's statutory body of authority so that it could accomplish what Congress had set it out to do in 1887: stabilize the railroad cartels without pooling. The rejection of rate increases between 1910 and the 5-percent case of 1914, with which the final third of the book is mainly occupied, can also be used to demonstrate how greatly the legislation of this period facilitated joint action by the railroads. Had the cartels not been stabilized by the legislation of 1903, 1906, and 1910, it is questionable whether the railroads could have mounted a collective effort to increase rates.

The denial of these increases must be considered in the context of a cartelized industry, and one which was on the eve of its decline. The purpose of the collusion in the industry was to administer a discriminatory pricing structure, to charge on the basis of the value of the service rather than the cost of the service. This left the industry open to "cream-skimming" by a carrier which could provide a higher quality of service for the freight high in value relative to weight against which the discrimination was directed. This carrier was to come in the form of the truck; but the first major effort at a land-borne rival to the railroads was already in existence in the form of the interurbans. Mainly built between 1899 and 1908, the interurbans were essentially an adaptation of streetcar technology for movement of short-distance inter-city passengers, express, and less-than-carload freight. Trucks and buses were available from about 1914.

Had the ICC behaved as Martin wants, by granting the proposed rate increases of 1910–1914, it would have accomplished approximately what it did in the early 1920s: armed with powers of minimum rate regulation and a target rate of return for the industry set in the Transportation Act of 1920, it pursued a policy of aggressive rate increases, thereby stimulating inter-city truck transportation, which became a major rival to the railroads by 1926. Similar policies in the 1910–1914 period would have accelerated this development by some years. Had the Commission accelerated the rate of investment in the industry by allowing rate increases in 1910–1914, it would have contributed to the excessive volume of capital characteristic of cartelized industries, and added to the problem of disinvestment once the industry began to decline.

The decline of the railroads was inevitable, given their industrial organization and technology. The ICC contributed to the decline not by denying a spirit of enterprise, but rather by enforcing the discriminatory pricing structure which the rise of competitive carriers would have swept away. This pricing structure made it impossible for the carriers to realize their comparative advantages for carriage of various distances, degrees of damage-proneness, and urgency. The railroads were severely handicapped in efforts to reattract long-distance traffic from trucks and bulk cargoes from barges by rate reduction. In a later period, the ICC contributed to the decline of the railroads by being strongly biased toward existing rail technology of individual cars. The Commission inhibited the development of integral-train technology for barge-competitive traffic by showing great reluctance to accept multi-car rates on bulk commodities, and in 1931 truncated the development of containerized technology for truck-competitive traffic, both in the interest of protection of existing rate structures based on carload lots. Car-service and safety requirements are also based on an existing technology of separate cars, which entails low productivity of the equipment and an excessive volume of damage claims. Thus, regulation is heavily committed to a technology which is failing a market test. This situation is likely to force a major change in the regulatory framework, or to bring further federal participation in the industry in the form of providing car supply or outright nationalization.

In concluding Martin addresses himself to present policy. Since his book is in a sense hostile to the Commission's historical behavior, one can expect him to be unsympathetic to its present behavior. Consistent with his interpretation of the Commission's basic nature he opposes outright abolition of the Commission [p. 365]:

> There remains today, even more emphatically than in 1906, the need for an ombudsman agency to rule on the reasonableness of individual rates; to prevent discrimination between persons and between places, whether due to ignorance, arrogance, bureaucratic callousness, or just plain dishonesty, and to eliminate rates which are patently too high in relation to the value of the service. Such regulation by the Commission, or some similar body, can and should continue, even if we should be so courageous and intelligent as to remove the fetters of minimum rates and the hobble of advance approval of rate changes.

Here again, Martin shows no recognition that he is dealing with a cartelized industry. The "discrimination between persons" which he still wishes prohibited is deviation from collusively set rates. The actual discrimination in the industry, in the economist's sense, is the value-of-service rate making which the entire edifice of regulation is intended to preserve. The passage just quoted appears to indicate that the author accepts value of service as the appropriate base of rate making even today; thus he is consistent in wishing to see the general framework of regulation retained. Significantly, he does not advocate repeal of the industry's antitrust exemption.

Intellectually, this book is a good demonstration of the familiar proposition of economists that the alternative to orthodox economic theory is not nontheoretical writing, but rather the development of casual or erroneous theory, leading to inappropriate and undesirable conclusions. Furthermore, the book well demonstrates the consequences of violating the historian's rule, "Study problems, not periods." In spite of its intellectual shortcomings, the work is assured a wide audience and, in particular, a favorable reception in the railroad industry, since it tells railroad executives what many of them want to hear by way of historical interpretation, and provides justification for the partial deregulation which most of them advocate, but without actual decartelization.

The determination of the allowed rate of return in a formal regulatory hearing

Paul L. Joskow
Assistant Professor of Economics
Massachusetts Institute of Technology

This paper presents an attempt to specify and estimate a simple model of the rate of return phase of a formal regulatory hearing. The specification of the model derives from the author's intensive study of the regulatory decision-making process in New York State. Although well-defined legal rules for the calculation of the allowed rate of return have not evolved in most regulatory jurisdictions, the results indicate that the regulatory agency, in a consistent fashion, makes use of the information provided to it in the regulatory hearing. The rate of return allowed by the commission is shown to depend on the size and relative reasonableness of the firm's request, the presence or absence of cost of capital testimony supporting the firm's request, the presence or absence of intervenors presenting conflicting rate of return testimony, the type of firm making the rate of return request, and a subjective evaluation of the efficiency of the firm making the request. The results also suggest that commission behavior changes in response to problems faced by the regulatory process in a world characterized by rapid inflation.

1. Introduction

■ Recent theoretical literature has analyzed the behavior of a neoclassical profit maximizing monopoly firm subject to a fair rate of return constraint, where the fair rate of return lies somewhere between the profit maximizing rate of return and the market cost of capital.[1] Although attempts have been made to derive the "optimal" fair rate of return in the context of this neoclassical model of the regulated firm,[2] no attempt has been made to analyze how regulatory commissions actually arrive at their allowed rate of return determination in a formal regulatory hearing.

This paper develops a simple model of the rate of return phase of a formal regulatory hearing in an effort to determine what factors affect the rate of return set by the regulatory authority. The model presented is based in part on the author's case study of the regulation of the gas and electric industries in New York by the New York State Public Service Commission.[3]

The author would like to thank Alvin Klevorick, John McGowan, and Richard R. Nelson for reading earlier drafts of this paper and for providing helpful comments and encouragement. Discussions with Joseph Swidler and W. K. Jones of the New York State Public Service Commission provided useful insights into the regulatory process in New York State. Remaining errors are those of the author.

[1] See Averch and Johnson [1] and Baumol and Klevorick [2].

[2] See Klevorick [5] and Sheshinski [6]

[3] See Joskow [4], Chapter 2, "Regulation of the Gas and Electric Industries in New York State: 1960–1970."

The formal hearing structure that has evolved for setting public utility prices is generally similar from one state to the next. The utility commission first seeks to determine the elements of the firm's capital stock that will be included in the firm's rate base. The commission then attempts to determine which elements of test year costs and revenues are to be allowed for regulatory purposes and whether or not specific changes which have occurred since the test year will be taken into consideration. Finally, the commission must determine what the fair rate of return is for the company.[4]

Legal rules specifying the items which may and may not be included in the rate base and in the allowable costs and revenues used for regulatory purposes have evolved in most regulatory jurisdictions. Similar legal rules specifying the proper method for calculating the allowed rate of return have not evolved, however. Commissions recognize that there are many things which must be taken into account in determining the allowed rate of return,[5] and they generally claim that they have fair, rational, and consistent methods for using the information at their disposal to arrive at an allowed rate of return in a regulatory proceeding. If this is in fact true we should be able to observe, specify, and estimate decision equations for the rate of return phase of a regulatory hearing.

2. The model of the rate of return phase

■ In this section a two-equation model of the rate of return phase of the formal hearing process is specified and estimated. The first equation is a commission decision equation, representing the allowed rate of return determination of the commission. The second equation is a firm request equation representing the request made by a firm in a given proceeding.[6]

In making its decision, the commission must attempt to set a rate of return that is not "so low" that the firm cannot perform its service function adequately (or even remain in business) and not "so high" that the firm is being allowed earnings above the amount needed to enable it to maintain the desired level of service quality. If the state public utility commission possessed the ability to define and understand the relationship between observable capital costs (debt costs), unobservable capital costs (the cost of equity capital), and the (unobservable) overall market cost of capital to the company, it could compute the market cost of capital faced by the firm. Futher, if the commission had a clear notion of the tradeoffs between the return allowed by the commission and the ability of the firm to achieve the desired service objectives, its job would be easy: it would simply calculate the market cost of capital faced by the firm and set a rule that the allowed rate of return should be some amount greater.

---

[4] A fourth and potentially important aspect of the formal hearing process deals with the question of rate structure. Most states have only recently begun to tackle the question of appropriate rate structures and their computation.

[5] See New York State Public Service Commission Opinion in Case 19757, *Re Consolidated Edison Company of New York,* for a discussion of the large number of factors which the regulatory commission takes into consideration in making the rate of return determination.

[6] A separate intervenor recommendation equation, symmetric to the firm request equation, is not estimated in this analysis, since the sample of data used here has intervenors presenting rate of return testimony in less than half of the cases As a result, we are interested more in the impact of the presence or absence of intervenors than in how the intervenor arrives at his recommendation.

No such calculations are made by most state regulatory agencies. There is a great deal of controversy, even among the experts who present rate of return testimony, over what the market cost of capital actually is. Moreover, it is doubtful that more than a handful of state regulatory commissioners have the ability to critically scrutinize the sometimes complex analyses employed by cost of capital witnesses, let alone make their own cost of capital calculations. In addition, commissions appear to possess only the most rudimentary understanding of the relationship between the return a firm is permitted to earn and the operational objectives that the commission wishes to achieve.

The ability of the commission to scientifically evaluate the rate of return requests made by the firms is therefore probably quite limited. As a result it appears that commissions rely on the general feeling that the return requested by the firm is probably too high on historical relationships between rate of return requests and observable capital costs and on information presented by intervenors.

3. The commission decision equation

■ The commission arrives at its conclusion about the fair rate of return by examining the presentation of the parties, applying its own judgment concerning their relative merits, and possibly making certain adjustments because of the performance characteristics of the particular firm. Equation (1) represents this relationship:

$$R^* = H(P_f, P_I, J, E, U_1), \tag{1}$$

where

U_1 = random disturbance term,
R^* = allowed rate of return,
P_f = presentation of firm in the hearing,
P_I = presentation of intervenors,
J = judgment of the commission, and
E = performance characteristics of the firm: subjective judgment on the part of the commission concerning the efficiency of the particular firm (the commission may want to reward firms that performed well).

The functional form and the magnitude of the associated coefficients reflect the relative weights the commission puts on the different presentations, and in this way they embody in part the "informed judgment" of the commission.

In specifying the commission decision equation the explanatory variables will be represented in the following way.

P_f: The presentation of the firm is represented by two variables:
R_f = the allowed rate of return requested by the firm. This variable embodies the firm's assessment of its cost of capital.
T = a dummy variable that equals unity if the firm presents testimony on the cost of equity capital and zero if it

does not. This variable is included for two reasons: (1) The presentation of testimony on the cost of equity bolsters the credibility of the firm's rate of return request and gives the commission more information on which to base its decision. (2) Certain commissioners in New York State have expressed displeasure with firms that have not included such testimony.[7]

P_I: The presentation of the intervenor. If the intervenor had presented rate of return testimony in each proceeding, similar variables could be included for the intervenor. Intervenors, however, do not always appear. In order to capture the impact of the existence of intervenor testimony as well as the difference between the intervenor recommendation and the company request (a measure of the extent of conflict) the following variable (I) was constructed.

I: Takes on the value zero when there is no intervenor presenting rate of return testimony and it equals $(1 - R_I/R_f)$ otherwise. The lower the value of the intervenor rate of return (R_I) relative to the requested rate of return of the firm, the higher the value of this variable, and vice versa. This variable will always lie between zero and one. It has been constructed in place of a simple dummy variable indicating the presence or absence of intervenors in an effort to capture the extent of conflict between the two parties.

J: The judgment of the commission. It is difficult to quantify the judgment of the commission concerning the reasonableness of the firm's presentation. Two variables are included here to represent the judgment of the commission.

- *GED*: A dummy variable, equal to one for gas rate cases and zero for electric cases, is used to investigate whether or not gas cases are treated differently from electric rate cases. It has been suggested elsewhere[8] that gas companies generally add a risk premium to their request, and since the commission has tended to reject the argument that gas companies are more risky, we would expect that gas company requests would be discounted more than electric company requests.[9]
- $\hat{U}_2$: The positive residuals calculated from the firm request equation (2) (which will be discussed presently) will be included. Equation (2) can be viewed as tracing out the average relationship between capital costs, company characteristics, and firm requests. Firms that stray above this average may be characterized as being relatively aggressive, trying to get more than the "typical" firm would request *ceteris paribus*. This variable is included to reflect not only the fact that the commission is concerned with

---

[7] See Dissent by Commissioner W. K. Jones in New York State Public Service Commission Case 25449, *Re Niagara Mohawk Company*.

[8] Joskow [4], pp. 42–43.

[9] "Special regulatory techniques, such as the purchased gas adjustment and weather normalization studies, tend to stabilize earnings." *Re Iriquois Gas Company*, 74 PUR 3d 348.

which party made the rate of return recommendation (firm or intervenor) but also that it has some rudimentary notion of the appropriate relationship between observable capital costs and "reasonable" rate of return requests. We would expect the commission to discount the requests of the relatively aggressive firms.

E: The performance potential is also very difficult to measure. It seems that those companies the commission views as efficient tend to avoid the formal hearing process.[10] As a result, there are not many observations on these companies. A dummy variable (E) that equals unity when the commission has commended a company for its efficiency or high quality of service and that equals zero otherwise was used to indicate the appearance of these firms.

In summary, the following variables are included in the commission decision equation:

R_c = rate of return allowed by commission,
R_f = rate of return requested by company,
T = dummy variable representing presence of cost of equity testimony,
I = variable representing the presence or absence of an intervenor and the degree of conflict between intervenor and firm where there is an intervenor presenting rate of return testimony,
GED = dummy variable (gas = 1, electric = 0),
$\hat{U}_2$ = positive residuals calculated from equation (2), and
E = dummy variable (efficient firms = 1, otherwise = 0).

A simple linear functional form has been chosen for estimation on the available data:[11]

$$R_c = b_0 + b_1 R_f + b_2 T + b_3 I + b_4 GED + b_5 \hat{U}_2 + b_6 E + U_1. \quad (1)$$

4. The firm request equation

■ The request of the firm is based on a cost of capital study that it presents and adjustments that it makes in the cost of capital calculation to take account of certain characteristics of the firm or the department of the firm involved in the hearing:

$$R_f = G_1(C_1, C_2, CAR_1, u_2), \quad (2)$$

---

[10] This was the case for most of the 1960s, the period from which the sample used for estimation is drawn. By 1970, rapid inflation was forcing all firms to the commission for price increases.

[11] Various nonlinear forms of the R_c equation were also estimated: quadratic in R_f, logarithmic in R_f and R_c, and logarithmic in R_f alone. There were no theoretical reasons for choosing a particular functional form and there was little ground for choice between the linear form and the nonlinear forms on goodness of fit or significance grounds. Introduction of nonlinear forms did not appear to improve the R^2 or change the signs of the coefficients or their classical significance levels appreciably. The analysis proceeds with the linear form, since this form is the easiest to interpret and exposit.

where

C_1 = imbedded cost of debt,
C_2 = firm's estimate of the cost of equity capital,
CAR_1 = relevant adjustments for particular characteristics of the firm (gas department, outside ownership, etc.), and
u_2 = random disturbance term.

Equation (2) represents the "average" request firms make. Variable C_1 is, for all intents and purposes, already determined and is not a center of controversy. Controversy arises over the assessment of the cost of equity capital; the differences over the assessment of this value constitute the major source of conflict between the firm's request and the intervenor's recommendation.

The firm request equation should in principle be a fairly simple one to estimate. After the presentation of its cost of capital testimony the company may make an upward or downward adjustment from this figure for its final rate of return request. It is hypothesized that this adjustment takes one of two forms:

(1) Gas departments of combination companies make an explicit or implicit upward adjustment in their requests relative to the calculated cost of capital, to reflect the notion that gas departments are "riskier" than electric departments.[12]

(2) There are some gas companies in the state that do not possess their own capital structures. These companies are entirely owned by corporations outside the state that are engaged in other activities besides gas distribution (retail) and electricity generation. It is suggested that these companies make a downward adjustment from their calculated capital costs since the capital structures used in their cost of capital presentation is that of their parent companies, which are riskier enterprises with high levels of equity.[13] The calculated cost of capital therefore includes an unjustified risk premium relative to the enterprise under consideration in the hearing.

One problem remains in specifying this equation for estimation. The firm's assessment of the cost of equity that we observe already includes, in many cases, these adjustments for firm characteristics. The firm request then reduces to a simple identity:

$$R_f = \alpha C_1 + \beta C_2^*,$$

where α and β are the proportions of debt and equity in the firm's capital structure, C_1 is the imbedded cost of debt, and C_2^* is the estimated cost of equity capital *adjusted for these risk characteristics.* To test the hypothesis that different kinds of firms discount their capital costs differently in requesting a rate of return we would like observations on the overall calculated cost of equity for the firm (C_2) before discounting for the particular subdivision of the firm making the price increase request. The request equation (2) would then be the following in linear form:

$$R_f = \delta_1 + \delta_2 C_1 + \delta_3 C_2 + \delta_4 CAR_1 + u_2.$$

---

[12] See Joskow [4], pp. 61–62.

[13] *Ibid.*

Data on the firm's overall calculated cost of equity are not available in general from the proceedings and it is treated as an omitted variable in this analysis where the omitted variable is assumed to be uncorrelated with the variables remaining.[14] The request equation to be estimated is therefore based on observable debt costs, and specified in linear form it becomes:[15]

$$R_f = a_0 + a_1 C_1 + a_2 GED_1 + a_3 OOD + u_2, \qquad (2)$$

where

C_1 = the imbedded cost of debt,

GED_1 = dummy variable taking on the value unity when the request is from a firm that is a gas department of a of a combination company and zero otherwise, and

OOD = dummy variable taking on the value unity when the request is from a company without its own capital structure and zero otherwise.

Equations (1) and (2) make up the behavioral model of the rate of return phase. This model is a two-equation recursive system:

$$R_c = b_0 + b_1 R_f + b_2 T + b_3 I + b_4 GED + b_5 \hat{U}_2 + b_6 E + u_1 \qquad (1)$$

$$R_f = a_0 + a_1 C_1 + a_2 GED + a_3 OOD + u_2 \qquad (2)$$

$$E(u_1 u_2') = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} E(u_1 u_1') = \sigma_1^2 I \quad E(u_2 u_2') = \sigma_2^2 I$$

I = identity matrix.

5. *A priori* discussion of the coefficients

■ The previous discussion leads to the expectation of the following signs for the firm request equation (2).

$(C_1)\hat{a}_1$: coefficient $\hat{a}_1$ should be positive. The requested rate of return should be positively associated with the cost of debt.

---

[14] This is a fairly strong assumption. One would expect that debt and equity costs would be positively correlated for a particular firm over time, reflecting movements in the general level of nominal interest rates and the changing riskiness of the firm. They would also be positively correlated across firms at any point in time, reflecting the relative riskiness of the firms Variable C_1 is not the cost of debt faced by a firm at any point in time, however, but a weighted average of the coupon rates of all of the firm's outstanding bond issues. As a result, one would expect the correlation between C_1 and C_2 to be smaller than the correlation between the cost of equity and the cost of debt faced by the firm at a point in time. To the extent that C_1 and C_2 are positively correlated, and since δ_3 is positive, estimating the request equation in linear form with C_2 omitted should result in an estimate of δ_2 [a_1 in the estimated equation (2)] that is biased upward.

[15] Three nonlinear functional forms were also estimated: quadratic in C_1 with a linear term, quadratic in C_1 without a linear term, and logarithmic in C_1. The quadratic form with the linear term had to be eliminated because of a collinearity problem between the linear term and the quadratic term. There was little ground for choice between the linear form and the remaining nonlinear forms on goodness of fit or significance grounds Introduction of nonlinear forms did not appear to improve the R^2 or change the signs of the coefficients or their classical significance levels appreciably. Although there may be some justification for preferring a nonlinear form of equation (2), the linear form is reported and discussed here primarily on the grounds of simplicity of interpretation and exposition.

$(GED_1)\hat{a}_2$: coefficient $\hat{a}_2$ should be positive. It reflects the risk premium gas departments of combination companies give to themselves.[16]

$(OOD)\hat{a}_3$: coefficient $\hat{a}_3$ should be negative. If the cost of capital testimony based on the capital structure of a more risky parent company is to be considered at all by the commission, the firm must request a lower rate of return than would be indicated by the cost of capital calculation.

The previous discussion leads one to expect certain signs and magnitudes for the coefficients of the linear commission decision equation (1):

$(R_f)\hat{b}_1$: The coefficient $b_1 < 1$ implies that the commission increases the allowed rate of return by less than the increase in the company's request. We would expect b_1 to be between zero and unity, since we do not expect the commission to allow the full amount of a firm's incremental request:

$$0 < b_1 < 1.$$

$(T)\hat{b}_2$: The sign of this coefficient should be positive, since the cost of capital testimony represents more evidence in support of the company request and because some commissioners have expressed their desire for such testimony.

$(I)\hat{b}_3$: The sign of this coefficient should be negative. Intervention should lower the allowed rate of return. If it does not, many people have been wasting time and money.

$(GED)\hat{b}_4$: The sign of this coefficient should be negative, reflecting the commission's attitude that the rates of return requested by gas companies have been unreasonably increased due to inclusion of an erroneous risk premium.

$(\hat{U}_2)\hat{b}_5$: The sign of this coefficient should be negative. Firms that make requests that are relatively "aggressive" should have their requests cut more than firms on or below this boundary.

$(E)\hat{b}_6$: The sign of this coefficient should be positive if the commission gives higher rates of return to companies with good records and should be zero if it treats these companies as it does all others.[17]

---

[16] If firms know that the regulators discount this premium it might be expected that groping toward equilibrium would suggest that the firms would omit it. This would be true unless (1) the commission does not fully discount the premium, or (2) the firms think that they can get the commission to change its procedure and grant risk premiums to gas departments

[17] The New York State Public Service Commission has indicated that firms which it recognizes as being exceptionally efficient may be rewarded for their superior performance. See Commission Opinion in Case 25393, *Re Long Island Lighting Company*, p. 13.

6. The data

■ The data come from twenty rate cases involving major gas or electric companies decided by the New York State Public Service Commission during the period 1960–1970. Where a variable was presented in the hearings as a range an average of the extrema was used. (For example, if the allowed rate of return was 6.0 to 6.4, the value 6.2 was used.) In two recent cases the commission staff called on an outside witness to present rate of return testimony. Since the appearance of expert witnesses for the staff appears to represent the commission assuming a new role as advocate, this testimony has been treated as intervenor testimony.[18]

7. Empirical results

■ Both equations of the model were fitted by ordinary least squares. The results for equation (2) are reported first.

Equation (2)

$$R_f = a_0 + a_1 C_1 + a_2 GED_1 + a_3 OOD + u_2$$

Coefficient	Estimated Value	t Statistic
$\hat{a}_0$	4.32	7.78
$\hat{a}_1$	0.64	5.17
$\hat{a}_2$	0.44	2.33
$\hat{a}_3$	−0.33	−1.64

$R^2 = 0.6468 \qquad F(3,16) = 9.7686$

The estimated coefficients all have the expected signs and all except $\hat{a}_3$ are significant at the 5-percent level. The coefficients $\hat{a}_2$ and $\hat{a}_3$ indicate the adjustments gas departments of combination companies and gas companies with outside capital structures make relative to their calculated historic cost of debt.

The estimates indicate that, *ceteris paribus*, an increase of one percentage point in the imbedded cost of debt will lead to an increase of 0.64 percentage points in the firm's request. Gas departments of combination companies will tend to add a premium of 0.44 percentage points to their requests. Firms with outside capital structures will tend to discount their request by 0.33 percentage points relative to other types of firms with the same calculated cost of capital.

Equation (1)

$$R_c = b_0 + b_1 R_c + b_2 T + b_3 I + b_4 GED + b_5 \hat{U}_2 + b_6 E + u_1$$

Coefficient	Estimated Value	t Statistic
$\hat{b}_0$	−1.14	−1.70
$\hat{b}_1$	1.10	10.85
$\hat{b}_2$	0.40	3.84
$\hat{b}_3$	−0.40	−3.89
$\hat{b}_4$	−0.20	2.18
$\hat{b}_5$	−0.64	−2.39
$\hat{b}_6$	0.11	1.05

$R^2 = 0.9449 \qquad F(6,13) = 37.1202$

---

[18] See *Annual Report of the New York State Public Service Commission, 1969*, p 5.

All coefficients have the expected sign and all of them except the coefficient of the efficiency variable (E) are significant at the 5-percent level. The coefficient b_1 is greater than unity and is highly significant, while the constant term is negative although only significant at the 10-percent level. The fact that $b_1 > 1$ is somewhat troublesome and will be discussed presently.

Coefficient b_2 is positive as expected. It indicates that the presence of cost of capital testimony is worth on the average 0.40 percentage points in the allowed rate of return.

Coefficient b_3 is negative as expected. It indicates that the presence of an intervenor will vary from no effect to a reduction of 0.40 percentage points in the allowed rate of return, depending on the degree of conflict between the petitioner and the intervenor. The average value of the intervenor variable is 0.43 and at this value intervention reduces the allowed rate of return by 0.172 percentage points.

Coefficient b_4 is also negative. The commission apparently does discount the requests of gas companies more than electric companies. This is probably a result of the effort by the commission to eliminate unjustified risk premiums which gas companies include in their requests. It appears, however, that gas departments may gain net since the average premium is 0.44 percentage points and the average discount by the commission is 0.20 percentage points: $b_1(0.44) - 0.20 = +0.35$ percentage points net. This may suggest why the gas departments continue to include this premium.

The fact that coefficient b_5 is negative is especially interesting. This indicates that aggressive firms that request more, *ceteris paribus*, than the "typical" firms will also have their requests cut more. A coefficient value of minus unity would indicate that these firms had the "aggressive" portions of their requests eliminated completely. Although b_5 has an absolute value less than unity, it is not significantly different from unity. This gives some indication that the commission feels that "legitimate" firm requests must bear some fairly constant relationship with the observable capital costs [C_1 in equation (2)].

Finally, b_6 is not significantly different from zero, but it does have the expected positive sign. This result gives some weak support to the hypothesis that the so-called highly efficient firms are treated somewhat better by the commission once they enter the formal hearing process.

What is troublesome about these results is the size of the coefficient of the firm request variable. One would expect *a priori* that the coefficient should be less than or equal to unity. The following "inflation premium" hypothesis might explain why the coefficient of R_f is greater than unity in these results. In an effort to make up losses that result from regulatory lag in an inflationary world the commission may allow a higher proportion of "corrected" firm requests during periods of rapid inflation than it would in a period of relatively stable prices.[19] For the period covered by the data the high rates of inflation are associated with increased capital costs and hence, via the request equation, with higher rate of return requests by the firms.

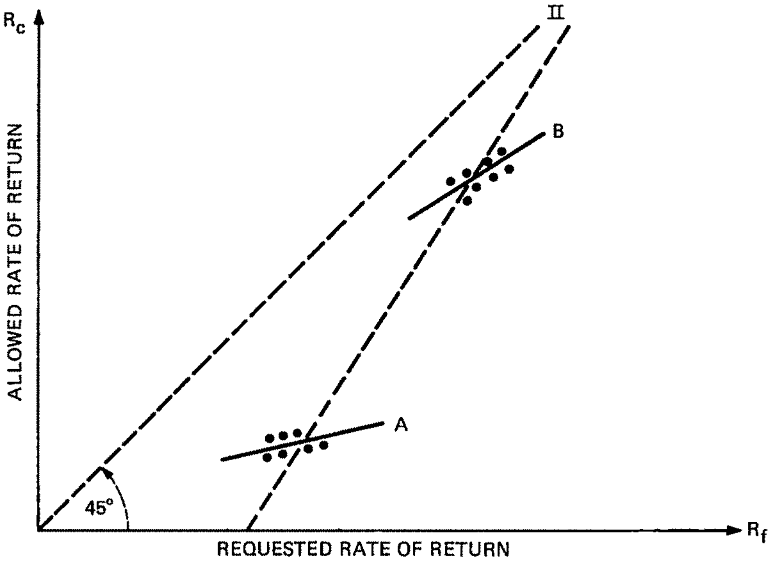
---

[19] Firms must normally continue to operate at the old prices while a hearing is taking place to determine whether a price increase is justified. Although most regulatory commissions have the power to allow proposed increases to go through pending the outcome of the hearing and then have the firm refund any excess, many commissions did not make extensive use of this power until 1971, when inflationary pressures became acute.

This situation is depicted in Figure 1. Line *A* represents the commission decision equation in R_c, R_f space for periods of stable prices, while line *B* represents the decision equation for periods of abnormally high inflation. The slope of *A* is less than *B* and the slope of both is less than unity. Fitting a regression line to the pooled data without taking into account the premium the commission grants for inflation would lead to a regression result as II in Figure 1, since the high inflation period is also the period of high R_f. The calculated regression line gets tilted upward with a slope that may be greater than unity. Inflation was a serious problem from about 1967 to the present. Beginning in 1968, decisions were based on test years for which inflation was becoming a noticeable problem.

FIGURE 1

RELATIONSHIP OF REQUESTED TO ALLOWED RATE OF RETURN UNDER DIFFERENT INFLATIONARY CONDITIONS



Ideally, to test this hypothesis we would like to split the sample between the high inflation years and the rest of the years and then estimate (1) separately for both sets of data. Unfortunately, with only 20 observations this is not possible. The following less rigorous attempts were made to cast light on the "inflation premium" hypothesis.

To conserve degrees of freedom, the variables representing particular characteristics of the hearing, *GED*, *T*, *I*, and *E*, were omitted and the following simple model of the commission decision equation was examined:

$$R_c = b_0 + b_1 R_f + b_2 \hat{U}_2 + u_1. \tag{1a}$$

Equation (1a) was first estimated using the full sample. Then, with the sample divided between (1) decisions reached before June 30, 1968 and (2) decisions reached after June 30, 1968, when the commission was apparently aware of the problem inflation was beginning to cause for regulated firms, (1a) was estimated for each subset.[20]

If the inflation premium hypothesis suggested above is correct, the estimates of b_0, b_1, and b_2 for the full sample should be very similar to the values obtained for the complete model:[21]

$$b_0 < 0$$

$$b_1 > 1$$

$$b_2 \approx -0.6.$$

For each of the grouped observations we would expect to observe the following signs and magnitudes for the regression coefficients:

Group 1: Noninflationary Period (*N*)
  Coefficient b_0^N should move in a positive direction from the result for full sample.
  Coefficient b_1^N should be less than unity.
  Coefficient b_2^N should be between zero and minus one.
Group 2: High Inflation (commission recognizes problem of inflation)
  Coefficient b_0^I should move in a positive direction from the result for the full sample.
  Coefficient b_1^I should be less than unity, but closer to unity than group 1.
  Coefficient b_2^I should be between zero and minus one.

In addition, we expect $b_0 < b_0^I < b_0^N$ (see Figure 1).

The results of the regression analysis are as follows:

Full Sample: 20 Observations $R_c = -1.09 + 1.10\,R_f - 0.81\,\hat{U}_2$
t values (-1.27) (9.04) (-2.21)
$R^2 = 0.84$ $SSR = 0.97.$

Group 1: 10 Observations $R_c = 1.87 + 0.65\,R_f - 0.56\,\hat{U}_2$
t values (1.14) (2.70) (-0.88)
$R^2 = 0.56$ $SSR = 0.26.$

Group 2: 10 Observations $R_c = 0.50 + 0.90\,R_f - 0.80\,\hat{U}_2$
t values (0.281) (3.79) (-1.81)
$R^2 = 0.67$ $SSR = 0.42.$

The regression results give some support to the "inflation premium" hypothesis. For the full sample the coefficient estimates are close to the corresponding ones for the full model. When the sample is divided we observe the behavior predicted by the hypothesis. The coefficient of R_f is less than unity for each subsample.

---

[20] Splitting the sample in this way is somewhat arbitrary. By the latter half of 1968 Commission Decisions appeared to take note of the fact that the rapid inflation, reflected in 1967 test year data and adjustments to it, was making it difficult for firms to even earn the returns that they had been allowed. The problem of firms returning for rate increase requests almost as soon as an increase has been granted did not become evident until 1969 and 1970.

[21] Equation (1a) is estimated for the entire sample to check that the omission of the other variables (*GED*, *T*, *I*, and *E*) does not effect the general nature of coefficients b_0, b_1, and b_2.

The value of the coefficient for the second group (inflationary) is closer to unity than for the first group (noninflationary). The constant terms move in a positive direction and have the predicted relative magnitude. Coefficient b_2 stays within the expected range. However, the application of an analysis of variance test (Chow test) does not allow us to reject the null hypothesis that the relationships for the subsamples are the same, at the 5-percent level.

8. Conclusions

■ The model of the rate of return phase presented indicates that although capital costs are a signficant basic component of the allowed rate of return, exactly what the commission will allow is determined by the size and relative reasonableness of the firm's request, the presence or absence of intervenors, the type of firm making the request, and the judgment of the commission. The commission behaves in a consistent fashion. There is no set legal rule for how the rate of return should be calculated, but once the firm's request is made, the type of firm determined, the presence of intervenors established, etc., we can predict, using equation (1), what the allowed rate of return will be.

In addition, the rate of return phase may be used to give compensation to firms for losses incurred because of regulatory lag and inflation. In an inflationary world, higher than normal rates of return may be allowed as a method of anticipating inflation and to forestall the immediate return of a firm for a rate increase.

Finally, the results indicate that statistical techniques can be of help in evaluating legal decision processes. Decision rules for legal decision-making units (whether they be regulatory agencies, other types of administrative agencies, or courts) should be observable, may be specified in a simple way, and can be estimated by collecting data from a sample of cases and using standard statistical techniques.[22] Such statistical analysis should give us a better idea of how various types of legal decision processes operate and provide valuable information for possible policy reform.

References

1. Averch, H and Johnson, H. L. "Behavior of the Firm under Regulatory Constraint" *The American Economic Review,* Vol 52, No 5 (December 1962), pp. 1062–67.
2. Baumol, W. J. and Klevorick, A. K. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science,* Vol. 1, No. 2 (Autumn 1970), pp. 162–90.
3. Davis, O., Dempster, M, and Wildavsky, A "A Theory of the Budgetary Process." *American Political Science Review,* Vol. 60, No. 3 (September 1966), pp. 529–47
4. Joskow, P. L. *A Behavioral Theory of Public Utility Regulation* Unpublished Ph.D. dissertation, Yale University, 1972
5. Klevorick, A. K "The 'Optimal' Fair Rate of Return." *The Bell Journal of Economics and Management Science,* Vol. 2, No 1 (Spring 1971), pp. 122–53.
6. Sheshinski, E. "Welfare Aspects of a Regulatory Constraint· Note." *The American Economic Review,* Vol 61, No. 1 (March 1971), pp. 175–78.

---

[22] Statistical analysis has been used in an analogous way by researchers investigating the federal budgetary process. See Davis, Dempster, and Wildavsky [3].

Investment behavior in the U. S. electric utility industry, 1949–1968

U. Sankar

Associate Professor of Economics
University of Wisconsin at Milwaukee

Jorgenson and Handel (J-H) applied their recently developed econometric model of investment behavior in U. S. regulated industries to four subindustries of the regulated sector. In their derivation of the demand for capital, they assumed constant returns to scale and unitary elasticity of substitution between capital and labor. For application to the U. S. electric utility industry the author finds it desirable to relax these two assumptions. Econometric studies on production and cost functions pertaining to this industry show that economies of scale are quite important and that the scope for substitution between capital and labor is limited. Using a CES production function, the author derives a more general expression for desired capital stock. This expression is incorporated in a logarithmic version of a rational distributed lag function in a manner suggested by Eisner and Nadiri. As in the J-H model, it is assumed that replacement investment is proportional to net capital stock.

The distributed lag function is applied to annual data relating to investor-owned electric utilities in the United States for the period 1949–1968. The estimates of long-run elasticities of capital stock with respect to relative price and output are of reasonable magnitudes and are in accord with results obtained from production function studies. The author finds that an investment function derived from a Leontief-type model is suitable for analyzing investment decisions in this industry. As for the characterization of the time structure of investment, he concludes that Jorgenson's rational distributed lag function is preferable to geometric distributed lag functions.

1. Introduction

■ Jorgenson and Handel (J-H)[1] have developed an econometric model of investment behavior for the regulated industries in the United States and have applied it to four subindustries of the regulated sector. As in the earlier contributions of Jorgenson and Jorgenson and Stephenson,[2] this model has three components: (1) demand for capital, (2) replacement investment, and (3) the relationship between changes in demand for capital and actual investment expenditures. J-H assert that hypotheses about the second and third components are the same for regulated industries as for unregulated industries. With respect to component (1) they postulate that for regulated industries the theory of optimal capital accumulation, which determines demand for capital, is the same as for unregu-

---

The author wishes to thank Professors Robert Eisner, Dale Jorgenson, and Arnold Zellner for many helpful comments, Lu Chu Chien and Ota Tatsuyuki for research assistance, and the American Telephone and Telegraph Co for financial support

[1] In [18]

[2] See [16] and [19], respectively.

lated industries, except that an additional constraint is imposed on the maximization of present value. Given the output price set by the regulator, the quantity of output is determined by the consuming public. Their marginal productivity condition for capital is

$$\frac{\partial Q}{\partial K} = \frac{c}{p + \pi},$$

where c is the rental price of capital services, p is the price of output, and π is "the difference between the price set by regulation and the price at which the level of output set by the consuming public $\bar{Q}$, under regulation, would be produced without regulation."[3] They make two crucial assumptions. First, they assume that regulation is effective—that is, that regulated price equals opportunity cost. They also observe that when regulation is effective, "the actual price and output are the same as those that would prevail under a competitive market structure."[4] Second, they assume that the production function has a Cobb-Douglas form. Utilizing these two assumptions, their marginal productivity condition for capital becomes

$$\alpha\frac{Q}{K} = \frac{c}{p},$$

where α is the elasticity of output with respect to capital. Thus the expression for desired level of capital,

$$K^{+} = \alpha\frac{pQ}{c},$$

is now of the same form for regulated industries as for unregulated industries.

In their concluding section, J-H suggest that it is desirable to test the hypothesis that the price set by regulatory agencies equals the opportunity cost of meeting demand for service. In some industries at least—notably in electric utilities—it is not possible by current regulatory policies to achieve the competitive ideal of setting marginal cost equal to price. The reason is that substantial economies of scale persist up to the largest sizes of units that have been built. Indeed, it is long-run decreasing cost that constitutes the justification for considering electric utilities as "natural monopolies."[5] There is also some econometric evidence[6] indicating that the scope for substitution between capital and labor is limited in electric power generation.

Jorgenson and Handel combine the effects of the relative price variable (p/c) and the output variable (Q) in one variable (pQ/c). In previous research on investment behavior in public utilities, investigators have used either output variables alone or relative price and output as separate variables. Chenery and Koyck[7] used output variables alone. Kisselgoff and Modigliani[8] experimented with different measures of input price variables and concluded that cost considerations appear to play a minor role in the investment decisions of electric utilities. Hickman[9] introduced both output and relative factor price as separate variables and reported that the coefficient of the relative factor price variable was not statistically significant (at the 5-percent level) from zero.

---

[3] [18], p 216

[4] *Ibid.*

[5] For an evaluation of the concept of natural monopoly and its importance in public utility economics, see Kahn [20], vol. 2, pp. 113–71.

[6] See Section 2.

[7] In [5] and [23], respectively.

Our objective in this paper is to analyze investment behavior in investor-owned electric utilities in the United States for the period 1949–1968. For application to privately owned electric utilities, we think it is worthwhile to relax two of the crucial assumptions of the J-H model and to investigate the separate effects of output and relative price variables on investment decisions. In Section 2 we review available empirical evidence on economies of scale and elasticities of substitution and suggest that a constant elasticity of substitution (CES) production function would provide a more realistic description of the technology. In Section 3 we derive an expression for desired level of capital stock which includes the J-H expression as a special case. We incorporate the desired capital stock function in a rational distributed lag function in a manner suggested by Eisner and Nadiri.[10] Following a stochastic specification, *a priori* restrictions on the coefficients of lag polynomials are given. Section 4 contains data sources and describes the procedures used in the construction of variables. Section 5 reports regression results. Estimates of elasticities and characteristics of lag distribution are obtained from regressions involving two different measures of the rental price variable. These results are compared with the results obtained from a naive model which includes output variables alone. In Section 6 we make some concluding remarks.

2. Evidence on returns to scale and capital-labor substitution in the electric utility industry

■ Econometric studies concerned with the estimation of returns to scale and factor substitution in the U. S. electric utility industry include the contributions of Komiya, Barzel, Nerlove, and Dhrymes and Kurz.[11] Komiya, Barzel, and Dhrymes and Kurz take plant as the unit of analysis and utilize Federal Power Commission data.[12] Nerlove's analysis is at firm level and he utilizes other FPC data.[13] Komiya and Barzel do not assume any optimizing behavior on the part of the entrepreneur. Nerlove and Dhrymes and Kurz assume cost minimization.

☐ **Economies of scale.** Komiya fitted Leontief-type production functions to 235 plants, newly built between 1938 and 1956. He classified plants into four groups according to the time period in which they were built, and further into coal and noncoal subgroups, according to the fuel used. He assumed each of three inputs —fuel, capital, and labor—as a logarithmic linear function of output. His estimates of elasticities of capital and fuel with respect to output range between 0.50 and 0.60. These results show that economies of scale are important in steam power generation.

---

[8] In [21].

[9] In [14].

[10] In [10]

[11] See [22], [2], [26], and [7], respectively.

[12] Given in *Steam Electric Plants · Construction Costs and Annual Production Expenses* [30].

[13] Given in *Statistics of Electric Utilities in the United States* [29].

Barzel used data for the 1941–1959 period and estimated demand functions for fuel, capital, and labor. For fuel and labor regressions he used information relating to all plants, old and new. For capital regressions, he included only observations on newly built plants. His regression equations include additional variables such as load factor, age of plant, input prices, and a set of dummy variables. His estimates of the partial elasticities of the quantity of fuel, labor, and capital with respect to output are 0.896, 0.626, and 0.815, respectively. All of these coefficients are significantly (at the 1-percent level) smaller than unity.

Nerlove suggested that a cost minimization model is appropriate for analyzing the behavior of regulated utilities. Assuming the Cobb-Douglas production function and competitive conditions in input markets, he derived a cost function. The reciprocal of the exponent of output in the cost function is a measure of returns to scale. Using data relating to a cross section of 145 firms in 1955, he obtained an estimate of 1.38 for returns to scale.[14]

Dhrymes and Kurz used a generalized CES function. Their production function permits the possibility that the (partial) elasticity of substitution between capital and fuel be different from the (partial) elasticity of substitution between capital and labor. They used data relating to newly constructed plants over the period 1937–1959 and stratified their sample by technological period and size of plant. Like Nerlove, they formulated a cost minimization model. Their estimates of returns to scale vary between 0.986 to 1.459.

All of these studies are concerned only with electricity generation in the steam power industry. Barzel[15] finds that economies due to the scale of the individual consumer (both residential and commercial) are also important in this industry.[16]

☐ **Elasticity of substitution between capital and labor.** Komiya fitted Cobb-Douglas production functions to his data and reported that

---

[14] Nerlove reports that the degree of returns to scale varies inversely with output. To investigate further the homotheticity property of the production function, we considered two returns-to-scale functions:

$$\mathrm{RTS} = \frac{1}{a + 2b \ln Q} \tag{i}$$

and

$$\mathrm{RTS} = \frac{A}{1 + BQ}, \tag{ii}$$

where Q is output, and a, b, A, and B are parameters The first function was used by Nerlove, and the second by Zellner and Revankar [31] Assuming the Cobb-Douglas production function and cost minimization, we estimated the returns-to-scale functions with data relating to 42 new steam plants (all single units) for the period 1961–1968. At unit level, we found that returns to scale increase with output. However, the relationships are not statistically significant at the 5-percent level Also, we incorporated the returns-to-scale functions in a Dhrymes-Kurz-type function (mixture of Leontief and neoclassical ingredients). Our results indicate that the relationships between returns to scale and output are not statistically significant.

[15] In [3]

[16] For other details pertaining to the technology of generating and transmitting electricity over the last two decades, see Hughes [15]

the results were unsatisfactory. Then he formulated and estimated Leontief-type production functions which do not permit substitution among inputs. Dhrymes and Kurz attempted to estimate the partial elasticity of substitution between capital and labor. They report that their results are disturbing for two reasons:[17]

> (a) the correlation coefficients are uniformly small and hence the explanatory power of the substitution model is very weak with respect to labor, (b) the coefficients, δ_1, which measure the sensitivity of labor input to changes in the relative prices, are essentially insignificant In only one out of thirteen cases are the δ_1 significantly different from zero at the 1-percent level of significance if a two-tailed T-test is employed.

They also provide an engineering explanation for the observed failure of the substitution model.

It should be noted that the studies of Komiya and Dhrymes and Kurz are concerned only with power generation in steam plants. Even here there is some scope for substitution between capital and labor in small plants.[18] Our study covers all electric operations: generation, transmission, and distribution. It is possible that the industry production function may allow substitution between capital and labor. For these reasons we prefer to work with a CES production function which allows the elasticity of substitution between capital and labor to take any value between zero and infinity. This function includes the Leontief-type production function (with zero elasticity of substitution) and the Cobb-Douglas function (with unitary elasticity of substitution) as special cases.

3. The model

■ **Demand for capital.** The production function for an electric utility,

$$Q = F(K,L), \tag{1}$$

where Q is output, K is capital, and L is labor, is assumed to satisfy the following conditions: (1) both inputs are necessary in producing output; (2) marginal physical products of both inputs are positive; and (3) the function F has continuous second-order partial derivatives with respect to capital and labor.

Following Nerlove, Dhrymes and Kurz, and Jorgenson and Handel,[19] we assume that the objective of the firm is to minimize cost for a given level of output. The equation for total cost is

$$C = wL + cK, \tag{2}$$

where w is the price of labor and c is the rental price of capital. The firm's problem is to minimize cost subject to the output constraint. The Lagrangian function for this problem is

$$\mathscr{L} = wL + cK + \lambda[\bar{Q} - F(K,L)]. \tag{3}$$

---

[17] [7], p 300.

[18] Dhrymes and Kurz report that, for small plants, the coefficient of δ_1 does in some cases exhibit statistical significance.

[19] In [26], [7], and [18], respectively The objective of the firm in the J-H model is to maximize its present value However, under their assumptions of given output and no adjustment costs, their equilibrium conditions are similar to the conditions in a cost-minimization model.

The first-order necessary conditions are:

$$w - \lambda F_L = 0 \tag{4}$$

$$c - \lambda F_K = 0 \tag{5}$$

$$\bar{Q} - F(K, L) = 0. \tag{6}$$

The second-order necessary condition is

$$\lambda(F_K{}^2 F_{LL} + F_L{}^2 F_{KK} - 2F_K F_L F_{KL}) < 0. \tag{7}$$

The Lagrangian multiplier, λ, may be interpreted as long-run marginal cost. Conditions (4) and (5) state that, in equilibrium, the product of long-run marginal cost and the marginal physical product must be equal to the input price. Condition (6) simply repeats the constraint. Since $\lambda > 0$, the condition that the expression in parentheses in equation (7) be negative is equivalent to the condition that isoquants are convex to the origin in the capital-labor space.

Let us consider the following CES production function:

$$Q = A[dK^{-\rho} + (1 - d)L^{-\rho}]^{-v/\rho},$$

where A is the (neutral) efficiency parameter, d is the distribution parameter, v is the scale parameter, and ρ is the substitution parameter.[20] The elasticity of substitution between capital and labor, σ, is $1/(1 + \rho)$.

For the CES production function the second-order necessary condition becomes

$$F_K{}^2 F_{LL} + F_L{}^2 F_{KK} - 2F_K F_L F_{KL} = -\frac{vQF_K F_L}{KL}(1 + \rho).$$

Thus the condition that isoquants are convex to the origin requires that $\rho > -1$ ($\sigma \geq 0$). The restrictions on other parameters are

$$A > 0, 0 < d < 1, \text{and } v > 0.$$

The necessary condition (5) may be written

$$c = \lambda A^{-\rho/v} d Q^{1+\rho/v} K^{-1-\rho} v. \tag{8}$$

We assume that the regulatory agency sets the price of output equal to long-run average cost (including a fair return on capital). For the CES function, long-run average cost is equal to the product of returns to scale (v) and long-run marginal cost (λ). Hence, we can write equation (8) as:

$$c = dpA^{-\rho/v} Q^{1+\rho/v} K^{-1-\rho}, \tag{9}$$

where p is the price of output. The expression for desired level of capital stock is[21]

---

[20] For an analysis of the properties of the CES function for the special case of $v = 1$, see Arrow et al. [1].

[21] It is possible to derive expressions identical to equation (10) under two other specifications about the technology. First, one may wish to ignore labor (since the value of labor is a small proportion of the value of output) and concentrate on the trade-off between fuel and capital. In that case one may specify a CES production function involving fuel and capital and interpret σ as the elasticity of substitution

$$K^{+} = (dA^{-\rho/v})^{1/(1+\rho)} Q^{(v+\rho)/v(1+\rho)} (p/c)^{1/(1+\rho)}$$

$$= \text{const}\,(p/c)^{\sigma} Q^{[\sigma + (1-\sigma)/v]}. \tag{10}$$

It may be noted that for the case of $\sigma = 1$, equation (10) is of the same form as that derived by Jorgenson and Handel. It may also be noted that for the case of $\sigma = 0$ equation (10) becomes

$$K^{+} = \text{const}\,Q^{1/v}, \tag{11}$$

which is of the form used by Koyck, Komiya, and Hickman.

☐ **Specification of the distributed lag function.** We assume that the relationship between changes in desired level of capital stock and actual changes in capital stock is of the following form:

$$\ln(K_{t+1}/K_t) = \mu(s)\ln(K_t^{+}/K_{t-1}^{+}), \tag{12}$$

where $\mu(s)$ is a polynomial in the lag operator. Under the assumption that the sequence of coefficients $\{\mu_\tau\}$ of the distributed lag function has a rational generating function,[22] equation (12) may be written

$$\Delta \ln K_t = \frac{\gamma(s)}{\omega(s)} \Delta \ln K_t^{+}, \tag{13}$$

where $\gamma(s)$ and $\omega(s)$ are polynomials in the lag operators.

For $\gamma(s)$ of order m and $\omega(s)$ of order n, after substituting equation (10) into equation (13) and adding an error term ϵ_t, the distributed lag function may be written

$$\Delta \ln K_t = \sum_{i=0}^{m} [\gamma_{pi} \ln(p/c)_{t-i} + \gamma_{qi} \Delta \ln Q_{t-i}]$$

$$- \sum_{j=1}^{n} \omega_j \Delta \ln K_{t-j} + \epsilon_t. \tag{14}$$

This equation is of the same form as the one used by Eisner and Nadiri.[23]

---

between fuel and capital. Second, one may wish to specify a CES production function involving capital, labor, and fuel. This specification implies that all partial elasticities of substitution are equal. Under any of these two specifications, equations (8), (9), and (10) are valid Our expression for desired capital stock does not involve the price of labor or the price of fuel. It is desirable to consider a more general functional form of the Dhrymes-Kurz type·

$$Q = [a_1 K^{-\rho_1} + a_2 L^{-\rho_2} + a_3 F^{-\rho_3}]^{-v/\rho}.$$

This would permit the partial elasticity of substitution between capital and fuel to be different from the partial elasticity of substitution between capital and labor. The marginal condition for capital is

$$c = \bar{\lambda} v(\rho_1/\rho) a_1 Q^{1+\rho/v} K^{-1-\rho_1}.$$

This equation resembles equation (8). But under the average cost pricing assumption it is not possible to find any simple relationship between $\bar{\lambda}$ (long-run marginal cost) and price of output. The difficulty arises because this production function is not homogeneous of any degree.

[22] In [17] Jorgenson introduced rational distributed lag functions. For further discussion, including a survey of the literature, see Griliches [13].

[23] In [10].

Following Jorgenson and Stephenson[24] we assume that the error term ϵ_t is distributed independently and identically over time and distributed independently of all values of changes in desired capital stock.

Jorgenson derives certain restrictions on the parameters of the lag polynomials.[25] For the special case of $\omega(s)$ of order 2, to satisfy the condition that the coefficients of the distributed lag function be nonnegative, we require that the leading coefficients of the $\gamma(s)$ polynomials be nonnegative, and that $-4\omega_2 \geq \omega_1{}^2$. The distributed lag function is a stable difference equation if the sum of ω_1 and ω_2 is less than one.

If the assumptions about the error term and the stability condition are satisfied, Jorgenson and Stephenson point out that the ordinary least squares estimator of the unknown parameters is best asymptotically normal.[26]

4. Data

■ The model developed in Section 3 is applied to annual data relating to investor-owned electric utilities in the United States. We obtained necessary data from the publications of the Edison Electric Institute (EEI) and of the FPC and from *Moody's Public Utilities*.[27] Regressions based on the lag specification in equation (14) have been fitted to data for the period 1949–1968. We also used the data for 1946–1948 in measuring lagged variables. A description of the measurement of variables used in regression analysis is given below.

☐ **Capital stock.** Under the assumption that replacement investment is proportional to net capital stock, it is possible to construct capital stock series if we are given (1) an estimate of the initial capital stock, (2) the depreciation rate, and (3) the series on gross investment.

First, we computed capital stock series in constant dollars using Ulmer's procedures. For gross investment in current prices for the period 1925–1968, we used the EEI data on construction expenditures in investor-owned electric utilities; for earlier years we used Ulmer's data.[28] These figures were reduced to constant (1954) prices by the use of a price index for investment goods. Using deflated gross investment data and Ulmer's assumptions about average life of plant and equipment, we constructed capital stock series.

Second, we used the estimates of capital stock for the years 1946 and 1969 and gross investment in 1954 dollars to obtain an estimate of the depreciation rate, δ. The estimated value of δ is 0.04843.

Finally, we computed the capital stock series (as of January 1) in 1954 dollars using the following formula:

$$K_{t+1} = I_t + (1 - \delta)K_t.$$

These series are given in Table 1.

---

[24] [19], p. 181.

[25] For details, see Jorgenson [17], pp. 146–47

[26] See [19], p 181

[27] See [8, 9], [29], and [25], respectively.

[28] [28], pp. 320–21.

TABLE 1

ELECTRIC LIGHT AND POWER (PRIVATE): VALUE OF CAPITAL STOCK, OUTPUT, PRICE OF OUTPUT, COST OF CAPITAL, CONSTRUCTION COST INDEX, AND RENTAL PRICE OF CAPITAL

YEAR	(1) VALUE OF CAPITAL STOCK (MILLIONS OF 1954 DOLLARS)	(2) OUTPUT KILOWATT-HOURS (BILLIONS)	(3) PRICE OF OUTPUT (CENTS)	(4) COST OF CAPITAL (PERCENT)	(5) CONSTRUCTION COST INDEX (1954 = 1.0)	(6) RENTAL PRICE OF CAPITAL (c_1)
1946	18,436	181	1.83	2.51	0.569	0.0418
1947	18,719	208	1.80	2.59	0.661	0.0491
1948	19,737	228	1.84	2.81	0.728	0.0557
1949	21,369	233	1.90	2.67	0.770	0.0579
1950	23,263	267	1.85	2.62	0.807	0.0602
1951	24,757	302	1.82	2.88	0.902	0.0697
1952	25,930	322	1.84	2.99	0.924	0.0724
1953	27,484	354	1.84	3.24	0.972	0.0786
1954	29,086	371	1.88	2.93	1.000	0.0777
1955	30,512	421	1.80	3.09	1.030	0.0817
1956	31,702	459	1.78	3.39	1.117	0.0920
1957	32,757	481	1.80	3.96	1.188	0.1046
1958	34,266	490	1.85	3.87	1.220	0.1063
1959	35,678	544	1.82	4.49	1.242	0.1159
1960	36,643	580	1.82	4.47	1.240	0.1155
1961	37,534	607	1.82	4.37	1.226	0.1130
1962	38,406	653	1.80	4.35	1.231	0.1132
1963	39,122	701	1.77	4.27	1.232	0.1123
1964	39,934	756	1.73	4.42	1.256	0.1163
1965	40,848	809	1.70	4.50	1.305	0.1219
1966	42,012	881	1.67	5.19	1.341	0.1345
1967	43,715•	928	1.66	5.58	1.390	0.1449
1968	46,036	1,019	1.64	6.22	1.439	0.1592
1969	48,848					

☐ **Output.** Our measure of output, Q, is electricity generated by investor-owned utilities (in billions of kilowatt-hours).[29]

☐ **Price of output.** Our measure of price of output, p, is average revenue per kilowatt-hour sold (in cents).[30]

☐ **Price index for investment goods.** We use the Handy-Whitman index (1954 = 1.0) as our measure of the price of investment goods, g. This index is a composite of over 50 materials and labor items included in the Handy-Whitman Index Service of cost indexes of construction and equipment: all steam.[31]

☐ **Cost of capital.** There are many difficulties in measuring the cost of capital appropriate for investment decisions. In a recent paper[32] Miller and Modigliani developed a measure of the cost of capital and obtained estimates for the electric utility industry for the years 1954–1957. They found that their estimates of the cost of capital conformed quite closely in their movements with the average yield on AAA bonds in the industry. For this reason, we use Moody's average yield on AAA bonds in public utilities as a surrogate for the cost of capital, r.

---

[29] See EEI [8, 9].

[30] *Ibid*

[31] *Ibid.*

[32] [24].

☐ **Rental price of capital.** We consider two measures of rental price of capital. The first measure, c_1, is equal to $(r_t + \delta)g_t$. The depreciation rate δ is taken as 0.04843.

Our second measure, c_2, is similar to the one used by Jorgenson and Handel. They characterize the U. S. tax structure by the corporation income tax rate, u, the proportion of depreciation for tax purposes to economic depreciation,[33] v, the proportion of the cost of capital deductible as interest, w, and the proportion of capital gains accrued but not realized, x. For this tax structure the rental price of capital becomes[34]

$$c_{2t} = g_t\left[\frac{1 - u_t v_t}{1 - u_t}\delta + \frac{1 - u_t w_t}{1 - u_t} r_t - \frac{1 - u_t x_t}{1 - u_t}\frac{\Delta g_t}{g_t}\right].$$

The tax rate u_t is taken to be equal to the marginal rate for corporate income. The proportion of depreciation v_t is the ratio of depreciation as given in the income statement to replacement in constant prices multiplied by the price of investment goods. The proportion of total cost of capital deductible for tax purposes is the ratio of long-term interest as given in the income statement to the total cost of capital, $r_t g_t K_t$. Following Jorgenson and Handel, we assume that all capital gains accrued or realized are regarded as transitory, so that the capital gains term $\Delta g_t/g_t$ may be taken to be identically zero.

The tax rate u_t, the proportions v_t and w_t, and the measure of rental price of capital c_{2t} are given in Table 2.

In a recent symposium,[35] Hall and Jorgenson, Bischoff, Coen, and Klein and Taubman suggested methods for incorporating changes in tax policies on the rental price of capital. In applying these methods to electric utilities we face two problems. First, with respect to depreciation policy, utilities are required to adopt the straight-line method for both book and rate-making purposes. For tax purposes, utilities may elect to follow double-declining or sum-of-the-years' digits methods. If they choose one of the accelerated depreciation methods, then the difference between the tax actually paid and the book tax can either be passed on to consumers in the form of a lower price (flow-through policy) or be set aside in a reserve for deferred taxes (normalization policy). The FPC first recommended normalization; since 1964 they have recommended the flow-through method. In a sample survey of electric utilities for the year 1965, Brigham[36] found that 17.2 percent of electric companies adopted straight-line, 20.4 percent sum-of-the-years' digits, and 62.4 percent double-declining balance. Of the companies electing to use accelerated depreciation, 54.5 percent normalize, while 45.5 percent

---

[33] Professor Eisner points out that the rental price variable ignores the corrections which were incorporated by Hall and Jorgenson in [13] and later offered by Coen in [6]. An appropriate measure of the rental price variable is

$$c_3 = g(\delta + r)(1 - uz)/(1 - u),$$

where z is the present value of the depreciation deduction on one dollar's investment

[34] See [18], p. 219.

[35] See Fromm [11]

[36] In [4].

TABLE 2

ELECTRIC LIGHT AND POWER (PRIVATE): DERIVATION OF JORGENSON'S MEASURE OF RENTAL PRICE OF CAPITAL, c_2*

$$c_{2t} = g_t\left[\frac{1-u_t v_t}{1-u_t}\,\delta_t + \frac{1-u_t w_t}{1-u_t}\,r_t\right]$$

YEAR	(1) u	(2) v	(3) w	(4) c_2
1946	0.3800	0.70861	0.72920	0.0491
1947	0.3800	0.62747	0.59600	0.0607
1948	0.3800	0.57913	0.52507	0.0708
1949	0.3800	0.54212	0.55312	0.0739
1950	0.4200	0.53124	0.52861	0.0807
1951	0.6775	0.48637	0.43537	0.1476
1952	0.7000	0.48434	0.43552	0.1626
1953	0.7000	0.47767	0.41246	0.1791
1954	0.5200	0.49126	0.46936	0.1213
1955	0.5200	0.50393	0.44485	0.1277
1956	0.5200	0.49272	0.38653	0.1469
1957	0.5200	0.48125	0.34003	0.1705
1958	0.5200	0.49096	0.37705	0.1707
1959	0.5200	0.50931	0.33574	0.1880
1960	0.5200	0.53714	0.35794	0.1841
1961	0.5200	0.57614	0.38738	0.1758
1962	0.5200	0.60489	0.40212	0.1734
1963	0.5200	0.63789	0.41787	0.1689
1964	0.5000	0.64839	0.40551	0.1707
1965	0.4800	0.64881	0.39520	0.1752
1966	0.4800	0.65019	0.35363	0.1970
1967	0.4800	0.64361	0.34595	0.2139
1968	0.4800	0.63398	0.33127	0.2380

* r AND g ARE GIVEN IN COLUMNS (4) AND (5) IN TABLE 1

flow through. Second, with respect to investment tax credits (generally about 3 percent of new depreciable assets), the impact of this credit can either be flowed through to reported profits in the year in which the investment is made or be amortized over a specified period. For the year 1965, Brigham reports that all the companies included in his sample took advantage of the credit. About 63 percent of these companies adopted normalization. Unfortunately, we do not have this type of information for other years.

It appears to us that the normalization procedure is equivalent to granting an interest-free loan to the utility, while the flow-through procedure is equivalent to reducing the price of output. To the extent that regulated agencies aim at maintaining a constant after-tax rate of return on the rate base, it seems reasonable to hypothesize that, apart from the liquidity effect, changes in tax policies affect primarily the price of output and not the rental price of capital. If this hypothesis is valid, then c_1 is an appropriate measure of the rental price of capital.

5. Results

■ Since we are dealing only with annual data, it is desirable to restrict the order of $\gamma(s)$ polynomial to three and the order of $\omega(s)$ polynomial to two. These restrictions would permit a maximum of three lagged changes in the relative price variable, three lagged changes in the output variable, and two lagged changes in the dependent variable. For annual data, these lag specifications are rich enough to include the lag specifications of Chenery, Koyck, Kisselgoff and Modigliani, and Hickman as special cases. Within our lag specifications it is possible to run several regression equations. Our procedure in the selection of the final regression is essentially the same as the one used by Jorgenson and Stephenson.[37] We run several regressions corresponding to different lag structures and choose the one which minimizes the standard error of the regression equation subject to the condition that the leading coefficients of the $\gamma(s)$ polynomials be nonnegative. If the fitted regression function does not satisfy the condition, $-4\omega_2 \geq \omega_1^2$, then the constraint is imposed by taking the unconstrained estimator of ω_1 as an estimator of ω_1 and ω_2 as $-\omega_1^2/4$. Estimators of the remaining coefficients will be obtained by rerunning the regressions with the parameters ω_1 and ω_2 set equal to their constrained estimates. Jorgenson and Stephenson point out that the resulting estimators of the regression coefficients are consistent but not necessarily efficient.

The order of presentation is as follows. First, we present regression results corresponding to two measures of rental price of capital and then analyze distributed lag responses due to changes in relative prices and output. Then we present results based on a Leontief-type production model. Finally, we compare the statistical fitting of the neoclassical model (corresponding to c_1) with that of the Leontief-type model.

☐ **Rental price of capital, c_1.** The regression function which minimizes the standard error of the regression (14) subject to the condition that the leading coefficients of the $\gamma(s)$ polynomials are nonnegative is[38]

$$\begin{aligned}\Delta \ln K_t = {} & \underset{(0.04473)}{0.04818}\, \Delta \ln(p/c_1)_{t-2} + \underset{(0.04824)}{0.18499}\, \Delta \ln Q_{t-2} \\ & + \underset{(0.21466)}{0.99306}\, \Delta \ln K_{t-1} - \underset{(0.17822)}{0.28984}\, \Delta \ln K_{t-2}\end{aligned}$$

$$s = 0.00710 \qquad R^2 = 0.83426 \qquad d = 1.54157.$$

The coefficient of multiple correlation, R^2, and the standard error, s, are corrected for degrees of freedom. The Durbin-Watson statistic, d, is biased toward randomness. This regression function does not satisfy the condition, $-\omega_2 \geq -\omega_1^2/4$. Hence, we re-estimated the regression using the unconstrained estimate of $\omega_1 = -0.99306$ and ω_2 as $-\omega_1^2/4 = 0.24654$. The resulting regression function is[39]

$$\begin{aligned}& \Delta \ln K_t - 0.99306\, \Delta \ln K_{t-1} + 0.24654\, \Delta \ln K_{t-2} \\ & \qquad = \underset{(0.03369)}{0.04643}\, \Delta \ln(p/c_1)_{t-2} + \underset{(0.03429)}{0.16368}\, \Delta \ln Q_{t-2}.\end{aligned}$$

---

[37] In [19].

[38] Equation (14) does not contain an intercept term. However, one can add an intercept and interpret it as the sample mean of the error in the equation. See Jorgenson and Handel [18], p. 221.

[39] The corresponding regression function including an intercept is

$$\begin{aligned}& \Delta \ln K_t - 0.99474\, \Delta \ln K_{t-1} + 0.24738\, \Delta \ln K_{t-2} \\ & \qquad = \underset{(0.00418)}{-0.00210} + \underset{(0.03663)}{0.05290}\, \Delta \ln(p/c_1)_{t-2} + \underset{(0.06573)}{0.19130}\, \Delta \ln Q_{t-2}.\end{aligned}$$

It may be noted that the coefficient of $\Delta \ln(p/c_1)_{t-2}$ is not significantly different from zero at the 5-percent level.

Estimates of long-run elasticities of capital stock with respect to relative price and output are:

$$\hat{E}_{(p/c_1)} = \hat{\sigma} = \frac{\hat{\gamma}_{p2}}{1 + \hat{\omega}_1 + \hat{\omega}_2} = 0.18317$$

$$\hat{E}_Q = \widehat{\left(\sigma + \frac{1-\sigma}{v}\right)} = \frac{\hat{\gamma}_{q2}}{1 + \hat{\omega}_1 + \hat{\omega}_2} = 0.64573.$$

The implied point estimate of the returns-to-scale parameter, v, is 1.76589.

☐ **Rental price of capital, c_2.** The regression function which minimizes the regression (14) (with c_2 as the rental price of capital) subject to the condition that the leading coefficients of the $\gamma(s)$ polynomials are nonnegative is

$$\begin{aligned}\Delta \ln K_t = {} & \underset{(0.01103)}{0.00945}\, \Delta \ln(p/c_2)_{t-2} + \underset{(0.04356)}{0.16784}\, \Delta \ln Q_{t-2} \\ & + \underset{(0.18463)}{0.90932}\, \Delta \ln K_{t-1} - \underset{(0.15857)}{0.23026}\, \Delta \ln K_{t-2}\end{aligned}$$

$$s = 0.00719 \qquad R^2 = 0.83003 \qquad d = 1.62234.$$

The corresponding constrained regression is

$$\begin{aligned}\Delta \ln K_t - 0.90932\, \Delta \ln K_{t-1} + 0.20672\, \Delta \ln K_{t-2} \\ = \underset{(0.00958)}{0.00884}\, \Delta \ln(p/c_2)_{t-2} + \underset{(0.02142)}{\Delta 0.15632}\, \ln Q_{t-2}.\end{aligned}$$

The estimated long-run elasticities of capital stock with respect to relative price and output are:

$$\hat{E}_{(p/c_2)} = 0.02972$$

and

$$\hat{E}_Q = 0.52562.$$

The estimate of $\hat{\sigma}$ is smaller now. The ratio of γ_{p2} to its standard error is so small that it is not possible to reject the null hypothesis that $\gamma_{p2} = 0$ (hence $\sigma = 0$) at any reasonable level of significance.[40]

☐ **Time structure of net investment.** The responses of net investment to relative price and output changes corresponding to the two measures of the rental price variable are given in Tables 3 and 4. It can be seen from columns (2) and (5) in these tables that the

---

[40] We also experimented with a third measure of rental price variable, c_3 (see note 33 above). The constrained regression function corresponding to this variable is

$$\begin{aligned}\Delta \ln K_t - 0.98271\, \Delta \ln K_{t-1} + 0.24143\, \Delta \ln K_{t-2} \\ = \underset{(0.02406)}{0.04180}\, \Delta \ln(p/c_3)_{t-2} + \underset{(0.03051)}{0.16969}\, \Delta \ln Q_{t-2}.\end{aligned}$$

The implied long-run elasticities of capital stock with respect to relative price and output are 0 16156 and 0.65588, respectively.

TABLE 3

DISTRIBUTED LAG COEFFICIENTS: RESPONSES OF INVESTMENT TO RELATIVE PRICE AND OUTPUT CHANGES CORRESPONDING TO THE RENTAL PRICE OF CAPITAL, c_1*

$$\Delta \ln K_t = 0.04643 \ \Delta \ln (p/c_l)_{t-2} + 0.16368 \ \Delta \ln Q_{t-2}$$

$$+0.99306 \ \Delta \ln K_{t-1} - 0.24654 \ \Delta \ln K_{t-2} + \hat{\epsilon}_t$$

$$= \sum_{i=1}^{\infty} [\hat{\mu}_{pi} \Delta \ln (p/c_l)_{t-i} + \hat{\mu}_{qi} \Delta \ln Q_{t-i}] + \hat{\epsilon}'_t$$

(1) i	(2) $\hat{\mu}_{pi}$	(3) $\sum_{k=1}^{i} \hat{\mu}_{pk}$	(4) $(3)/\hat{E}_p$	(5) $\hat{\mu}_{qi}$	(6) $\sum_{k=1}^{i} \hat{\mu}_{qk}$	(7) $(6)/\hat{E}_q$
1	0	0	0	0	0	
2	0.04643	0.04643	0.25348	0.16368	0.16368	0.25348
3	0.04611	0.09254	0.50521	0.16254	0.32622	0.50520
4	0.03434	0.12688	0.69269	0.12106	0.44728	0.69267
5	0.02273	0.14961	0.81678	0.08015	0.52743	0.81680
6	0.01410	0.16371	0.89376	0.04974	0.57717	0.89383
7	0.00840	0.17211	0.93962	0.02963	0.60680	0.93971
8	0.00486	0.17697	0.96615	0.01716	0.62396	0.96629
9	0.00276	0.17973	0.98122	0.00974	0.63370	0.98137
10	0.00154	0.18127	0.98963	0.00544	0.63914	0.98979
11 TO ∞	0.00190			0.00659		
SUM	0.18317		1.00000	0.64573		1.00000
MODAL LAG	2.00000			2.00000		
MEDIAN LAG	2.91000			2.91000		
MEAN LAG	3.97000			3.97000		

*SINCE WE HAVE ONLY ONE SIMILARLY LAGGED TERM FOR EACH OF THE PRICE AND OUTPUT VARIABLES, WE HAVE IN FACT CONSTRAINED COLUMNS (4) AND (7) TO BE EQUAL. THE TRIVIAL DIFFERENCES MUST BE DUE TO LACK OF PRECISENESS IN COMPUTATIONS.

TABLE 4

DISTRIBUTED LAG COEFFICIENTS: RESPONSES OF INVESTMENT TO RELATIVE PRICE AND OUTPUT CHANGES CORRESPONDING TO THE RENTAL PRICE OF CAPITAL, c_2*

$$\Delta \ln K_t = 0.00884 \ \Delta \ln (p/c_2)_{t-2} + 0.15632 \ \Delta \ln Q_{t-2}$$

$$+ 0.90932 \ \Delta \ln K_{t-1} - 0.20672 \ \Delta \ln K_{t-2} + \hat{\epsilon}_t$$

$$= \sum_{i=1}^{\infty} [\hat{\mu}_{pi} \Delta \ln (p/c_2)_{t-i} + \hat{\mu}_{qi} \Delta \ln Q_{t-2}] + \hat{\epsilon}'_t$$

(1) i	(2) $\hat{\mu}_{pi}$	(3) $\sum_{k=1}^{i} \hat{\mu}_{pk}$	(4) $(3)/\hat{E}_p$	(5) $\hat{\mu}_{qi}$	(6) $\sum_{k=1}^{i} \hat{\mu}_{qk}$	(7) $(6)/\hat{E}_q$
1	0	0	0	0	0	0
2	0.00884	0.00884	0.29744	0.15632	0.15632	0.29740
3	0.00804	0.01688	0.56797	0.14214	0.29846	0.56782
4	0.00548	0.02236	0.75236	0.09694	0.39540	0.75225
5	0.00332	0.02568	0.86406	0.05877	0.45417	0.86407
6	0.00189	0.02757	0.92766	0.03340	0.48757	0.92761
7	0.00102	0.02860	0.96231	0.01822	0.50579	0.96227
8	0.00054	0.02914	0.98048	0.00966	0.51545	0.98065
9	0.00028	0.02942	0.98991	0.00502	0.52047	0.99020
10	0.00014	0.02956	0.99462	0.00258	0.52304	0.99509
11 TO ∞	0.00160			0.00257		
SUM	0.02972		1.00000	0.52562		1.00000
MODAL LAG	2.00000			2.00000		
MEDIAN LAG	2.75000			2.75000		
MEAN LAG	3.66000			3.67000		

*SINCE WE HAVE ONLY ONE SIMILARLY LAGGED TERM FOR EACH OF THE PRICE AND OUTPUT VARIABLES, WE HAVE IN FACT CONSTRAINED COLUMNS (4) AND (7) TO BE EQUAL. THE TRIVIAL DIFFERENCES MUST BE DUE TO LACK OF PRECISENESS IN COMPUTATIONS.

relative weights are zero in the first period, reach the maximum in the second period, and then gradually decline thereafter. Our estimates of average lags are smaller than those of Chenery (4.00) and Koyck (13.5), but higher than that of Jorgenson and Handel (2.50). Columns (4) and (7) show that about 50 percent of responses to changes in output or relative price occur by the end of the third year, and more than 80 percent by the end of the fifth year.

☐ **Leontief-type limitational model.** Our regression results show that economies of scale are quite important in the electric utility industry. These results confirm econometric results based on production functions and engineering estimates. Our results also indicate that the hypothesis of zero elasticity of substitution between capital and labor cannot be rejected at any reasonable level of significance. We have noted earlier that even *ex ante* substitution between capital and labor is limited in this industry.

Our results suggest that a Leontief-type production model is appropriate for analyzing investment decisions in this industry. Estimation of investment functions based on this model would be desirable for two reasons. First, earlier investigators such as Koyck and Hickman estimated investment functions using annual data and their results favor a limitational model. Both Koyck and Hickman used geometric distributed lag functions. It would be interesting to use a rational distributed lag function, since it includes the geometric lag function as a special case. Second, it is often argued that, because of long-term planning and large capital outlays associated with installing power plants, investment decisions in electric utilities are based solely on demand considerations. For these reasons, it would be interesting to compare the statistical fitting of the neoclassical model with the fitting of the Leontief-type model.

We substitute equation (11) in equation (13) and obtain the following distributed lag function:

$$\Delta \ln K_t = \sum_{i=0}^{m} \gamma_i \Delta \ln Q_{t-i} - \sum_{j=1}^{n} \omega_j \Delta \ln K_{t-j} + e_t. \qquad (15)$$

As in the neoclassical model considered earlier, we restrict the order of $\gamma(s)$ polynomial to three and the order of $\omega(s)$ polynomial to two. Six regression functions, corresponding to different lag specifications, satisfy the constraint that the leading coefficient of $\gamma(s)$ polynomial be nonnegative. For these regressions, information pertaining to the standard error, R^2, Durbin-Watson statistic, and elasticity of capital with respect to output are given in Table 5. The best fitting equation with a geometric distributed lag function $(\gamma_1, \gamma_2, \omega_1)$[41] has a standard error of 0.00702. The estimated value of elasticity of capital with respect to output for this regression equation is 0.53047. This is smaller than Hickman's estimate of 0.6898 for public utilities. The regression function which minimizes the standard error of equation (15) corresponds to a lag specification: $\gamma_1, \gamma_2, \omega_1, \omega_2$. However, this equation does not satisfy the

---

[41] The modification of Koyck's flexible accelerator model to incorporate an additional lagged term in output is due to Hickman [14], pp. 320–41. The low value of the Durbin-Watson statistic is an indication of positive autocorrelation.

TABLE 5

STANDARD ERROR OF REGRESSION EQUATION (15), R^2, DURBIN–WATSON STATISTIC, AND ELASTICITY OF CAPITAL WITH RESPECT TO OUTPUT FOR ALTERNATIVE FORMS OF THE POLYNOMIALS $\gamma(s)$ AND $\omega(s)$

(1) $\gamma(s)$	(2) $\omega(s)$	(3) STANDARD ERROR	(4) R^2 CORRECTED FOR DEGREES OF FREEDOM	(5) DURBIN–WATSON STATISTIC	(6) ELASTICITY OF CAPITAL WITH RESPECT TO OUTPUT
1	1	0.00892	0.73855	1.42668	0.48607
1	1,2[a]	0 00658	0 85795	2.47041	0.52835
2	1	0.00723	0.82813	1.34082	0.50172
2	1,2	0.00714	0.83269	1 68812	0.48966
1,2	1	0.00702	0 83797	0.82036	0.53047
1,2	1,2[a]	0.00561	0 89658	1 40438	0.53805

[a] THE CONSTRAINT $-4\omega_2 \geq -\omega_1^2$ IS NOT SATISFIED.

constraint $-4\omega_2 \geq -\omega_1^2$. Imposing this constraint, we reestimated this regression function. The resulting function is

$$\Delta \ln K_t - 1.03719\, \Delta \ln K_{t-1} + 0.26894\, \Delta \ln K_{t-2} = \underset{(0.03218)}{0.06468\, \Delta \ln Q_{t-1}} + \underset{(0.03032)}{0.06183\, \Delta \ln Q_{t-2}}.$$

The elasticity of capital with respect to output is

$$\hat{E}_Q = \left(\frac{\hat{1}}{v}\right) = \frac{0.06468 + 0.06183}{1 - 1.03718 + 0.26894} = 0.54589.$$

The time pattern of response of net investment to a change in output is given in Table 6. The relative response reaches its peak in the second year and declines thereafter. About 83 percent of the response

TABLE 6

DISTRIBUTED LAG COEFFICIENTS. RESPONSES OF INVESTMENT TO OUTPUT CHANGES

$$\Delta \ln K_t = 0.06468\, \Delta \ln Q_{t-1} + 0.06183\, \Delta \ln Q_{t-2}$$
$$+ 1.03719\, \Delta \ln K_{t-1} - 0\ 26894\, \Delta \ln K_{t-2} + \hat{e}_t$$
$$= \sum_{i=1}^{\infty} \hat{\mu}_{qi}\, \Delta \ln Q_{t-i} + \hat{e}'_t$$

(1) i	(2) $\hat{\mu}_{qi}$	(3) $\sum_{k=1}^{i} \hat{\mu}_{qk}$	(4) $(3)/\hat{E}_q$
1	0.06468	0 06468	0.11849
2	0.12892	0.19360	0.35465
3	0.11631	0.30991	0.56772
4	0.08597	0.39588	0 72520
5	0.05789	0.45377	0.83125
6	0.03692	0.49069	0 89888
7	0.02272	0.51341	0.94050
8	0.01363	0.52704	0 96547
9	0.00803	0.53507	0 98018
10	0.00466	0.53973	0 98872
11 TO ∞	0 00616		
SUM	0 54589		1 00000
MODAL LAG	2.00000		
MEDIAN LAG	2.68000		
MEAN LAG	3.64000		

occurs by the end of the fifth year. For public utilities, Hickman estimated that about 91 percent of the response occurs by the end of the fifth year. The higher rate of adjustment in Hickman's model may be due to the fact that his data cover a broader industrial group.

☐ **Comparison of the neoclassical model corresponding to c_1 and the Leontief-type model.** First, we compare the time pattern of response in net investment to a change in output. To aid in visualizing the time pattern, we have plotted the elements of the $\mu(s)$ polynomial in Figures 1 and 2. In the neoclassical model, the first period response is zero while in the Leontief-type model the first period relative response is about 0.12. In both models, the relative response reaches its peak in the second year and declines thereafter. The average and median lags are smaller in the Leontief-type model.

FIGURE 1

THE TIME FORM OF LAGGED RESPONSE OF NET INVESTMENT TO CHANGE IN OUTPUT: NEOCLASSICAL MODEL WITH c_1

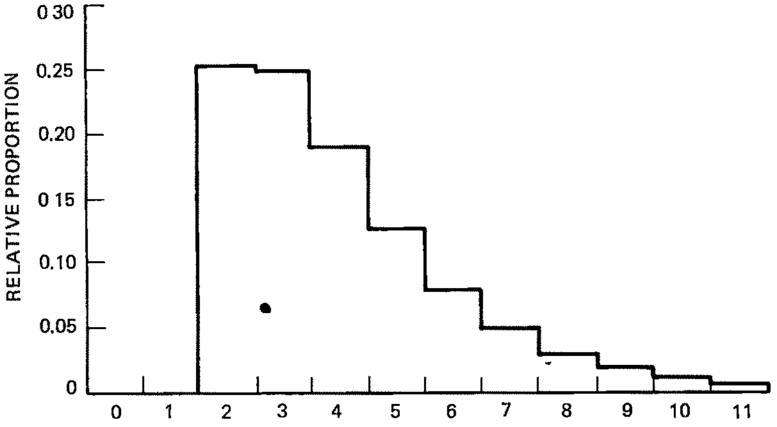
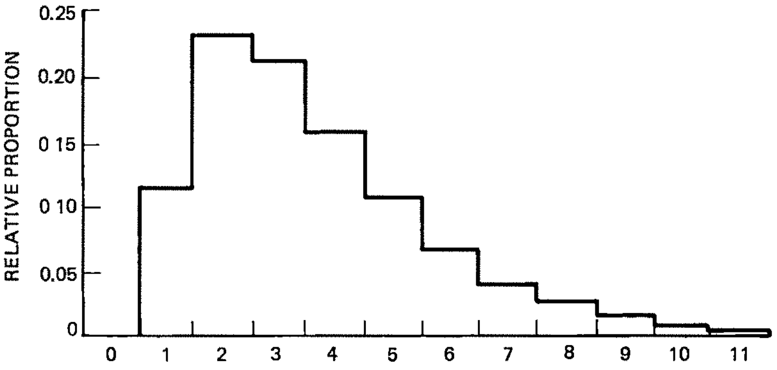


FIGURE 2

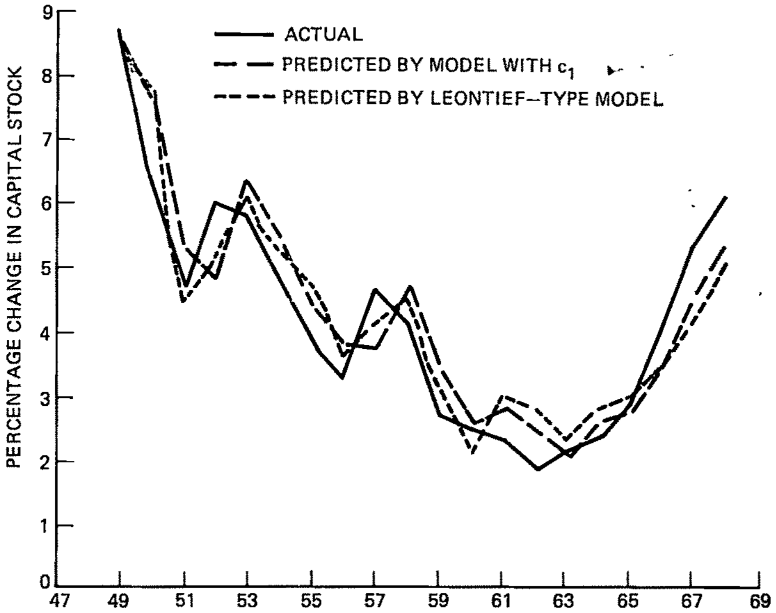
THE TIME FORM OF LAGGED RESPONSE OF NET INVESTMENT TO CHANGE IN OUTPUT: LEONTIEF–TYPE MODEL



Second, we compare the actual percentage change in capital with the predicted changes in both models. See Figure 3. The neoclassical model explains 88.68 percent of the variations in percentage change in capital stock; the corresponding percentage for the Leontief-type model is 92.14.

FIGURE 3

PERCENTAGE CHANGE IN CAPITAL STOCK, ACTUAL VS. PREDICTED



6. Concluding remarks

■ This paper modifies the Jorgenson-Handel model to allow for increasing returns to scale and to permit the elasticity of substitution between capital and labor to take any nonnegative value. This modification results in a logarithmic version of a rational distributed lag function, which is applied to data relating to investor-owned electric utilities in the United States.

Our method of constructing capital stock series is consistent with the assumption that replacement investment is proportional to net capital stock. We considered two different measures of the rental price of capital. For both measures, we estimated long-run elasticities of capital stock with respect to relative price and output and characterized lag distributions. Our estimates of the elasticities are of reasonable magnitudes and are in accord with the results based on production function studies.

We found that an investment function based on a Leontief-type production model performs slightly better than the functions based on the neoclassical model. This result confirms the commonly held view that, because of long-term planning considerations and large capital outlays associated with installing power plants, investment decisions in electric utilities are based solely on demand considerations.

Our estimates of the average lags are about 3.75 years. These estimates are considerably smaller than those of Chenery and Koyck, but slightly higher than that of Jorgenson and Handel. We note that the use of a geometric lag distribution results in a misspecification of the time structure of investment.

Our analysis is confined to privately owned electric utilities. Since there are substantial differences among regulated industries in technological characteristics, demand conditions, and regulatory policies, we think that analysis at the industry level is desirable. This would enable one to tap hitherto unused data sources and construct suitable measures of variables. There is one problem, however: detailed information is available only on an annual basis. For a fine characterization of the time structure of investment, quarterly data are desirable.

Reference

1 Arrow, K. J, Chenery, H. B., Minhas, B., and Solow, R M. "Capital-Labor Substitution and Economic Efficiency" *The Review of Economics and Statistics*, Vol. 43, No. 3 (August 1961), pp. 225–50.

2. Barzel, Y. "The Production Function and Technical Change in the Steam Power Industry." *Journal of Political Economy*, Vol. 72, No. 2 (March/April 1964), pp. 133–50.

3 ——— "Productivity in the Electric Power Industry, 1929–55." *The Review of Economics and Statistics*, Vol. 45, No. 4 (November 1963), pp. 395–408.

4. Brigham, E F. "Public Utility Depreciation Practices and Policies." *National Tax Journal*, Vol 19, No. 2 (June 1966), pp. 144–55.

5 Chenery, H. B "Overcapacity and the Acceleration Principle." *Econometrica*, Vol. 20, No. 1 (January 1952), pp 1–28

6 Coen, R. M "Effects of Tax Policy on Investment in Manufacturing" *The American Economic Review*, Vol 58, No. 2 (May 1968), pp. 200–11.

7. Dhrymes, P. J. and Kurz, M. "Technology and Scale in Electricity Generation." *Econometrica*, Vol. 32, No 3 (July 1964), pp. 287–315.

8. Edison Electric Institute *Historical Statistics of the Electric Utility Industry* 1963.

9. ——— *Statistical Yearbook*. 1969.

10. Eisner, R. and Nadiri, M. "Investment Behavior and Neo-Classical Theory" *The Review of Economics and Statistics*, Vol. 50, No. 3 (August 1968), pp 369–82

11. Fromm, G., ed. *Tax Incentives and Capital Spending* Washington, D C.. The Brookings Institution, 1971.

12. Griliches, Z. "Distributed Lags: A Survey." *Econometrica*, Vol. 35, No 1 (January 1967), pp. 16–49.

13. Hall, R. E and Jorgenson, D W. "Tax Policy and Investment Behavior." *The American Economic Review*, Vol. 57, No. 3 (June 1967), pp. 391–414.

14. Hickman, B. G *Investment Demand and U.S. Economic Growth*. Washington, D C.. The Brookings Institution, 1965.

15 Hughes, W. R. "Scale Frontiers in Electric Power," in W M. Capron, ed, *Technological Change in Regulated Industries*, Washington, D. C.: The Brookings Institution, 1971

16. Jorgenson, D. W. "Capital Theory and Investment Behavior." *The American Economic Review*, Vol. 53, No. 2 (May 1963), pp. 247–59.

17. ———. "Rational Distributed Lag Functions" *Econometrica*, Vol 34, No 1 (January 1966), pp 135–49.

18. ———, and Handel, S. S. "Investment Behavior in U. S. Regulated Industries" *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 213–64.

19. ———, and Stephenson, J. A. "Investment Behavior in U.S Manufacturing, 1947–60." *Econometrica*, Vol. 35, No 2 (April 1967), pp. 169–220.

20. Kahn, A E *The Economics of Regulation* 2 Vols New York· John Wiley & Sons, Inc., 1971.

21 Kisselgoff, A. and Modigliani, F. "Private Investment in the Electrical Power Industry and the Acceleration Principle," *The Review of Economics and Statistics*, Vol. 39, No 4 (November 1957), pp. 363–79

22. Komiya, R. "Technological Progress and the Production Function in the U. S. Steam Power Industry." *The Review of Economics and Statistics*, Vol 44, No. 2 (May 1962), pp. 156–66.

23. Koyck, L. M. *Distributed Lags and Investment Analysis*. Amsterdam· North-Holland Publishing Co, 1954.

24. Miller, M. H and Modigliani, F "Some Estimates of the Cost of Capital to the Electric Utility Industry." *The American Economic Review*, Vol. 56, No 3 (June 1966), pp 333–91.

25. Moody, John. *Moody's Public Utilities Manual*. New York: John Moody, 1971.

26. Nerlove, M. "Returns to Scale in Electricity Supply," in C Christ et al., eds., *Measurement in Economics.* Stanford, Calif.: Stanford Univ Press, 1963, pp. 167–98.
27. Phillips, C. F *The Economics of Regulation.* Homewood, Ill.: Richard D. Irwin, Inc, 1969.
28. Ulmer, M J. *Capital in Transportation, Communication, and Public Utilities* Princeton, N. J.: National Bureau of Economic Research, 1960
29. U S Federal Power Commission. *Statistics of Electric Utilities in the United States,* Classes A and B Privately Owned Companies. Annual Issues
30. ———. *Steam Electric Plants. Construction Cost and Annual Production Expenses.* Annual supplements
31. Zellner, A. and Revankar, N S. "Generalized Production Functions" *Review of Economic Studies,* Vol 36, No. 2 (April 1969), pp. 241–50.